REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	05 Aug 1994	Final 6 Jun 1993 5 Jun 1994

4. TITLE AND SUBTITLE
Joint Services Electronics Program
Final Report

5. FUNDING NUMBERS
F49620-93-C-0014

2305/AS
61102F

6. AUTHOR(S)
J. Bokor and M.A. Lieberman

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Electronics Research Laboratory
253 Cory Hall
University of California at Berkeley
Berkeley, CA 94720

8. PERFORMING ORGANIZATION REPORT NUMBER
UCB/ERL-93/1

AFOSR-TR- 95 0143

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NE
Directorate of Physics and Electronics
110 Duncan Avenue Suite B115
Bolling AFB DC 20332-0001
Program Manager: Lt.Col. Billy Smith

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
93-C-0014

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE.
DISTRIBUTION UNLIMITED.

12b. DISTRIBUTION CODE

Justification
By
Distribution/
Availability Codes

Dist	Avail and/or Special
A-1	

13. ABSTRACT *(Maximum 200 words)*

This report summarizes the research activity supported by the Joint Services Electronics Program at the Electronics Research Laboratory for the period June 1993 to June 1994. The Berkeley JSEP effort was organized into three themes: Quantum and Opto-electronic Devices, Semiconductor Electronic Devices, and Artificial Neural Networks. Under Quantum and Opto-electronic Devices, a new femtosecond laser laboratory for ultrafast measurements of carrier dynamics in semiconductors was completed. In the Semiconductor Electronic Device area, a novel device design was invented that has made possible quantitative measurement of the saturation velocity in inversion layers. This key parameter directly relates to the ultimate speed of transistors making it practical to realize CMOS-like circuits that can operate at voltages as low as 0.6V, and still maintain excellent speed and turn-off characteristics. Under Artificial Neural Networks, the focus continued to be the development of connectionist algorithms that are directly applicable to real-world problems, with particular attention given to applications of specific DoD interest. This work on speech recognition and machine vision emphasizes robustness in the face of noise.

14. SUBJECT TERMS	15. NUMBER OF PAGES
None	74
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	

SN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18

F. Assaderaghi, S. Parke, D. Sinitsky, P.K. Ko, and C. Hu, "Variable Threshold Voltage MOSFET (VTMOS) for Very Low Voltage Operation," submitted to the *Electron Device Letters Journal*, April 1994.

M. Chan, Z.-J. Ma, F. Assaderaghi, C.T. Nguyen, C. Hu and P.K. Ko, "A Low-Barrier Body-Contact Scheme for SOI MOSFETs to Eliminate the Floating Body Effect," *Proceedings, 1993 International Semiconductor Device Research Symposium*, December 1-3, 1993, pp. 341-344.

M. Chan, F. Assaderaghi, S.A. Parke, S.S. Yuen, C. Hu and P. K. Ko, "Recess Channel Structure for Reducing Source/Drain Series Resistance in Ultra-Thin SOI MOSFET," published in *Electronic Device Letter*, January, 1994

M. Chan, S. Yuen, Z.-J. Ma, K. Y. Hui, P. K. Ko and C. Hu, "Comparison of ESD Protection Capability of SOI and BULK CMOS Output Buffers," *1994 IEEE International Reliability Physics Proceedings*, pp. 292-298.

N.L. Chang and A. Zakhor, "Intermediate View Reconstruction for Three-Dimensional Scenes," *Proceedings, International Conference on Digital Signal Processing*, July 14-16, 1993, Vol. 2, pp. 636-641, Nicosia, Cyprus.

J. M. Cruz and L.O. Chua, "An IC Chip of Chua's Circuit," to appear in the *IEEE Transactions on Circuits and Systems, Part II: Analog and Digital Signal Processing*, Vol. 40, No. 10.

C. Hu, "Silicon-on-Inulator for High Speed ULSI," 1993 International Conference on Solid State Devices and Materials, Makuhari, 1993, pp. 137-139.

S. Im, Ph.D. Thesis, "Ion Beam Synthesis of SiGe Alloy Layers," submitted to the Graduate Division of the University of California at Berkeley, April 1994.

J. Koehler, N. Morgan, H. Hermansky, H. Guenter Hirsch, and G. Tong, "Integrating RASTA-PLP into Speech Recognition," accepted, in press for ICASSP, '94, Adelaide, Australia.

K. Lau, "Mode-Locked Semiconductor Lasers," Invited Talk, LEOS Annual Meeting, San Jose, 1993.

J. Malik and R. Rosenholtz, "A Differential Method for Computing Local Shape-From-Texture for Planar and Curved Surfaces," to appear in *Proceedings of the 27th Asilomar Conference on Signals, Systems and Computers*.

J. Malik and R. Rosenholtz, "Recovering Surface Curvature and Orientation From Texture Distortion: A Least Squares Algorithm and Sensitivity Analysis," presented at ECCV94, European Conference on Computer Vision '94, May, 1994, Stockholm, Sweden.

Z.J. Ma, H.J. Wann, M. Chan, J. King, Y.C. Cheng, P.K. Ko, and C. Hu, "Characterization of Hot-Carrier Effects in Thin-Film Fully-Depleted SOI MOSFETs," *1994 IEEE International Reliability Physics Proceedings*, pp. 52-56.

N. Morgan, "Current Research in Acoustically Robust Speech Recognition," *Proceedings of American Voice Input/Output Society*, (AVIOS), in Press.

Arlindo L. Oliveira and A. Sangiovanni-Vincentelli, "Learning Complex Boolean Functions : Algorithms and Applications", Proceedings of Neural Information Processing Systems '93, Denver, CO. Morgan Kaufmann.

G. Tong, M.S. Thesis, "Combating Additive Noise and Spectral Distortion in Speech Recognition Systems with JAH-RASTA," submitted to the Graduate Division of the University of California at Berkeley, May 1994.

P. Tsai, M.S. Thesis, "A Neural-Net Based, In-Line Focus/Exposure Monitor," Electrical Engineering Department, University of California at Berkeley, July 1994.

J.D. Walker, "Vertical-Cavity Laser Diodes Fabricated by Phase-Locked Epitaxy," Ph.D. Thesis, Electrical Engineering and Computer Science Department, University of California at Berkeley, May 1993.

H-J. Wann, J. King, J. Chen, P.K. Ko, and C. Hu, "Hot-Carrier Currents of SOI MOSFETs," pp. 118-119, *1993 IEEE International SOI Conference Proceedings*, October, 1993, Palm Springs, California.

E. LIST OF REPORTABLE INVENTIONS

- Title: Method of Producing Silicon-on-Insulator Substrates; Inventors: N. Cheung, C. Hu, J. Liu, and S.S.K. Iyer
- Title: A Dynamic Threshold Voltage MOSFET (DTMOS) for Ultra Low Voltage Operation; Inventors: C. Hu, P. Ko, F. Assaderaghi, S. Parke

F. APPENDIX OF JSEP-SPONSORED PUBLICATIONS

A Novel Silicon-On-Insulator (SOI) MOSFET for Ultra Low Voltage Operation

Fariborz Assaderaghi, Stephen Parke*, Ping K. Ko, and Chenming Hu
Department of Electrical Engineering and Computer Science
University of California at Berkeley, Berkeley, CA 94720
Phone: (510) 642-1010, FAX: (510) 642-2739, email: fariborz@diva.berkeley.edu
* IBM Corporation, East Fishkill, NY

To extend the lower bound of power supply voltage, we propose a Variable Threshold Voltage MOSFET (VTMOS) built on Silicon-On-Insulator (SOI). Threshold voltage of VTMOS drops as gate voltage is raised, resulting in a much higher current drive than regular MOSFET, at low V_{dd}. On the other hand, V_t is high at V_{gs}=0, thus the leakage current is low.

The SOI devices used in the study were built on SIMOX wafers. A four terminal layout was used to provide separate source, drain, gate, and body contacts. In addition to the four-terminal layout, devices with local gate-to-body connections were also fabricated as illustrated in Fig. 1. This connection uses an oversized metal to P+ contact window aligned over a "hole" in the poly gate [1]. The metal shorts the gate and P+ region. Thus, there is no significant penalty in area.

To operate the VTMOS, floating body and gate of a SOI MOSFET are tied together. This is not a new configuration, as [1-3] have already suggested it. However, [1-3] all tried to exploit the extra current produced by the lateral bipolar transistor. This normally requires the body voltage to be larger than 0.6V. Since current gain of the bipolar device is small, extra drain (collector) current comes at cost of excessive input (base) current, which contributes to the standby current. We will show that most of the improvement can be achieved when gate and body voltages are kept below 0.6V. This also ensures that base current will stay negligible. Although the same idea can be used in bulk devices, better advantage is reached in SOI, where because of very small junction areas base current and capacitances are appreciably reduced.

Fig. 2 illustrates the NMOS behavior, with a separate terminal used to control the body voltage. The threshold voltage at zero body bias is denoted by V_{to}. Body bias effect is normally studied in the reverse bias regime, where threshold voltage increases as body to source reverse bias is made larger. We propose to use the exact opposite regime. Namely, we "*forward bias*" the body-source junction (at less than 0.6V), forcing the threshold voltage to drop.

Specifically, this forward bias effect is achieved by connecting the gate to the body. This is shown as V_{gs}=V_{bs} line in Fig. 2. The intersect of V_t curve and V_{gs}=V_{bs} line, which is marked as V_{tf}, is the VTMOS threshold voltage. This lower threshold voltage does not come at expense of higher off-state leakage current, because at V_{bs}=V_{gs}=0 VTMOS and regular device have the same V_t. In fact, they are identical in all respects and consequently have the same leakage. This is clearly seen in Fig. 3. Reduced V_{tf} compared to V_{to} is attained through a theoretically ideal subthreshold swing of 60mV/dec. Fig. 3 demonstrates this for PMOS and NMOS devices operated in VTMOS mode and in regular mode.

This is not the only improvement. As the gate of VTMOS is raised above V_{tf}, threshold voltage drops further. For example, for tech-B in Fig. 2, at V_{gs}=V_{bs}=0.6V, V_t=0.18V compared to V_{to}=0.4V. In VTMOS operation the upper bound for applied V_{gs}=V_{bs} is set by the amount of base current that can be tolerated. This is illustrated in Fig. 3, where PMOS and NMOS device body (base) currents are shown. At V_{gs}=0.6V base currents for both PMOS and NMOS devices are less than 2nA/μm. Current drives of VTMOS and regular MOSFET are compared in Fig. 4, for tech-B of Fig. 2. VTMOS drain current is 2.5 times of regular device at V_{gs}=0.6V, and 5.5 times of regular device at V_{gs}=0.3V.

AC performance of VTMOS is evaluated by an unloaded 101 stage CMOS ring oscillator, shown in Fig. 5. We emphasize that since the threshold voltages of devices used in the ring oscillator were high (tech-A), the optimum performance was not achieved. For tech-B, ring oscillators are not available. If the devices based on tech-B are used, the expected delay for unloaded ring oscillator can be calculated by: $\tau_{pd} = \frac{C}{4} V_{dd}(\frac{1}{I_{dsatn}} + \frac{1}{I_{dsatp}})$. This is shown as solid squares in Fig. 5, where C=200fF is used for W_n=5μm and W_p=10μm. This value for C was obtained by fitting the equation to the measured τ_{pd} of tech-A. Fig. 6 illustrates the inverter DC transfer characteristics of tech-B.

Acknowledgment

This project was supported by SRC under Contract 93-DC-324, ISTO/SDIO through ONR under Contract N00014-92-J-1757, and AFOSR/JSEP under Contract F49620-93-C0041.

References

[1] S. A. Parke, C. Hu, and P. K. Ko, *IEEE Elect. Dev. Lett.*, vol. 14, no. 5, pp. 236-238, May 1993.

[2] J. P. Colinge, *IEEE Trans. Elect. Dev.*, vol. ED-34, no. 4, pp. 845-849, Apr. 1987.

[3] S. Verdonckt-Vandebroek, S. Wong, J. Woo, and P. Ko, *IEEE Trans. Elect. Dev.*, vol. 38, pp. 2487-2496, Nov. 1991.

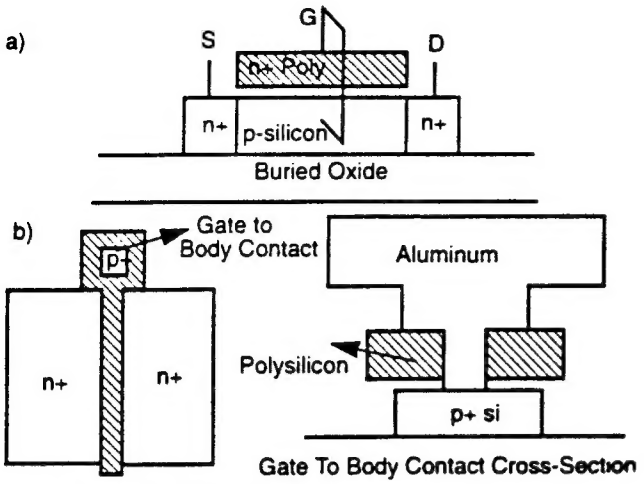


Fig. 1 **a)** Cross section of an SOI NMOSFET with body and gate tied together. **b)** Gate to body connection by using aluminum to short the gate and P+ region.

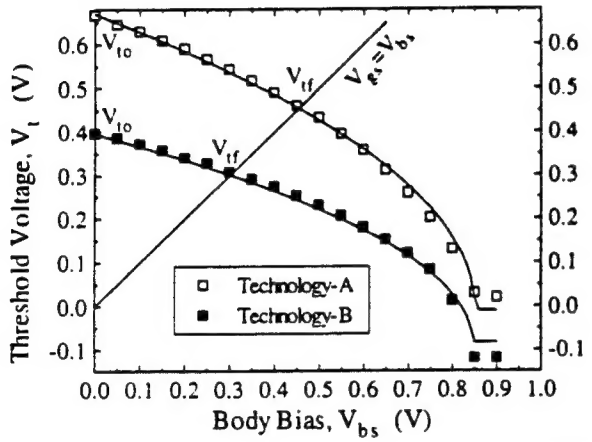


Fig. 2 Threshold Voltage of SOI NMOSFET as a function of body-source forward bias. For Tech-A T_{ox}=10nm, N_a=2.0x10^{17}cm^{-3}. For Tech-B T_{ox}=6.4nm N_a=2.3x10^{17}cm^{-3}.

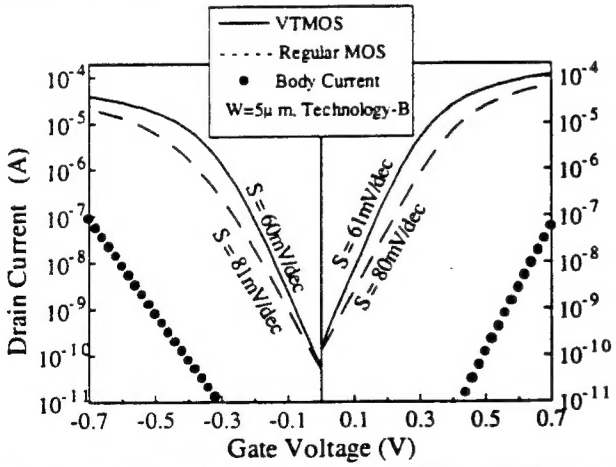


Fig. 3 Subthreshold characteristics of SOI NMOSFET and PMOSFET, with body grounded and body tied to the gate.

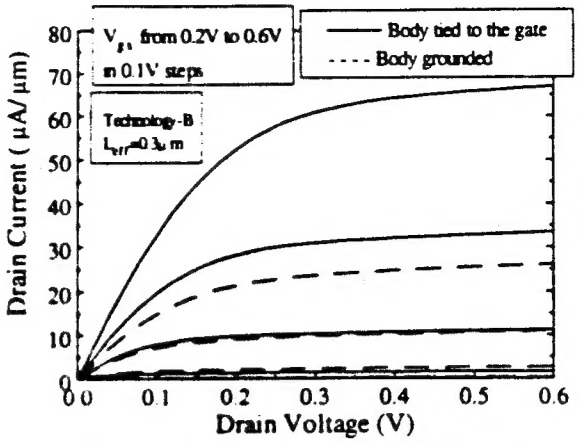


Fig. 4 Drain current of an SOI NMOSFET operated as a VTMOS and as a regular device.

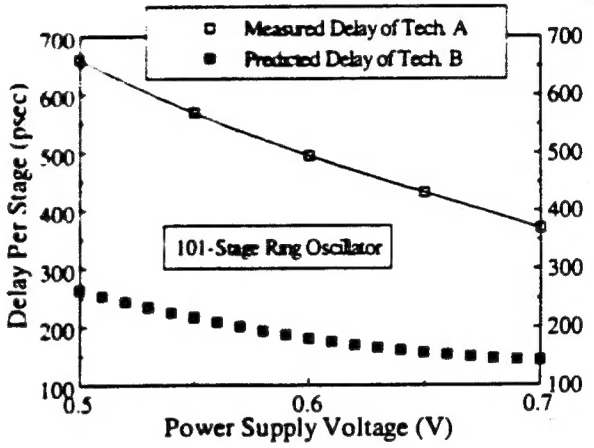


Fig. 5 Delay of a 101-stage ring oscillator. The PMOS and NMOS devices in the ring are VTMOS with T_{ox}=10nm, and L_{eff}=0.3μm, V_{to}=0.6V. Solid Squares show the predicted delay for a ring oscillator based on Tech-B with L_{eff}=0.3μm.

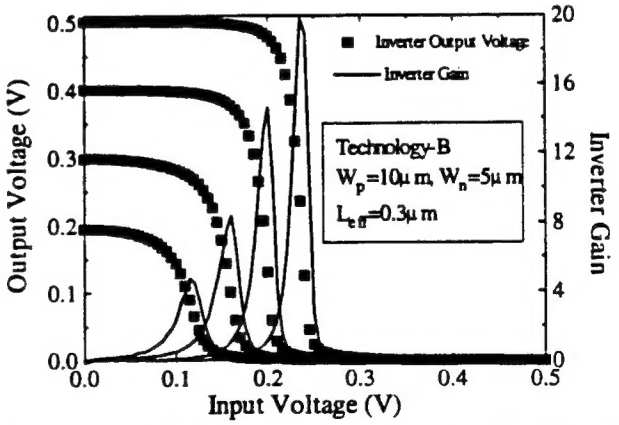


Fig. 6 Inverter DC transfer characterisctics. PMOS and NMOS devices forming the inverter are VTMOS.

High-Performance Sub-Quarter-Micrometer PMOSFET's on SOI

Fariborz Assaderaghi, Stephen Parke, Joe King, Jian Chen, *Member, IEEE*, Ping Keung Ko, *Member, IEEE*, and Chenming Hu, *Fellow, IEEE*

***Abstract*—PMOS transistors with effective channel lengths down to 0.15 μm have been fabricated on silicon-on-insulator (SOI) films. Gate oxide thicknesses of 5.5 and 10 nm are used. These P^+ gate PMOS devices exhibit excellent short-channel behavior, low source–drain resistance, and remarkably large current drive and transconductance. For T_{ox} = 5.5 nm, saturation transconductances of 274 mS/mm at 300 K and 352 mS/mm at 80 K are achieved, which are the highest reported values for this oxide thickness. The result is attributed to low series resistance, forward-bias body effect, and the reduction of body charge effect.**

I. Introduction

POTENTIAL advantages of MOS transistors built in thin SOI films include less process complexity, reduced parasitic capacitances, improved short channel effects, absence of latch-up, and higher transconductance and current drive. However, to date very few successful experimental results have been reported to substantiate improved current drive. Often high parasitic series resistance has obscured this advantage [1]. Here, for the first time, we report experimental results for deep-submicrometer SOI PMOSFET's with improved performance over their bulk counterparts.

II. Device Fabrication

A full description of the process integration is given in [2]. Here, we provide only the key processing steps. SIMOX substrates with a final SOI film thickness of 130 nm were used. The 130-nm film thickness permits low device series resistance without using silicidation. Also, by avoiding ultrathin films, desired threshold voltage can be easily achieved. Mesas were created by plasma etching a nitride/oxide/silicon stack stopping at buried oxide. Next a 100-nm oxide was grown on the mesa sidewalls to prevent low-V_t edge devices and gate oxide defects at the mesa corners. Threshold implants were then performed, resulting in concentrations of $1-3 \times 10^{17}$ cm^{-3}. Gate oxides of 5.5 and 10 nm thickness were grown, followed by the deposition of 280 nm of undoped polysilicon. Doping of the poly gate was realized by a 30-keV 5×10^{15} cm^{-2} boron implant. The combination of p^+-poly gate, silicon film thickness, and doping concentration resulted in nearly fully depleted (NFD) devices with threshold voltage range of -0.3 to -0.5 V. These threshold voltages are consistent with intended low-voltage operation of the devices. Effective channel lengths as short as 0.08 μm were obtained by O_2 ashing of the gate photoresist [3].

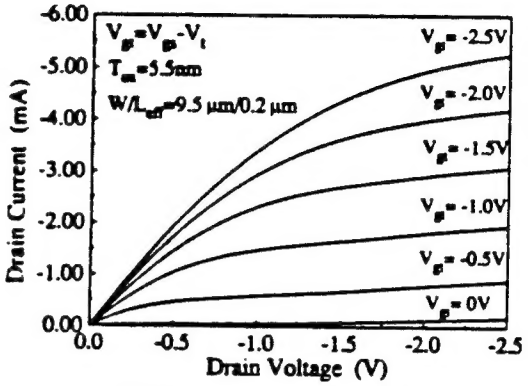


Fig. 1. *I–V* characteristics of a PMOSFET with W/L_{eff} = 9.5 μm/0.2 μm and T_{ox} = 5.5 nm.

III. Device Performance

We note that all reported channel lengths here are the effective channel lengths, determined from standard conductivity measurement, not the mask lengths. Fig. 1 shows the *I–V* characteristics of a 9.5-μm/0.2-μm device with T_{ox} = 5.5 nm. Although fabricated devices are NFD, and long-channel devices show kink in their *I–V*, very short devices have reduced kinks as seen in Fig. 1. This is due to the fact that the depletion regions of the source/drain junctions nearly deplete the film at drain voltages lower than the onset of the kink. Fig. 2 shows the subthreshold swing and threshold voltage shift (ΔV_t) of the fabricated PMOSFET's. Although ultrathin film is not used, good subthreshold swing and short-channel behavior is observed. Subthreshold characteristics of a device with L_{eff} = 0.2 μm are shown in Fig. 3.

Fig. 4 shows that for T_{ox} = 5.5 nm, the device with L_{eff} of 0.15 μm has saturation transconductances (G_m) of 274 mS/mm at room temperature and 352 mS/mm at 80 K. These are measured values and have not been corrected

Manuscript received January 19, 1993; revised April 6, 1993. This work was supported by the Semiconductor Research Corporation under Contract 93-DC-324, ISTO/SDIO through ONR under Contract N00014-92-J-1757, and AFOSR/JSEP under Contract F49620-90-C-0029.

F. Assaderaghi, S. Parke, J. King, P. K. Ko, and C. Hu are with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720.

J. Chen was with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720. He is now with Advanced Micro Devices, Sunnyvale, CA.

IEEE Log Number 9209605.

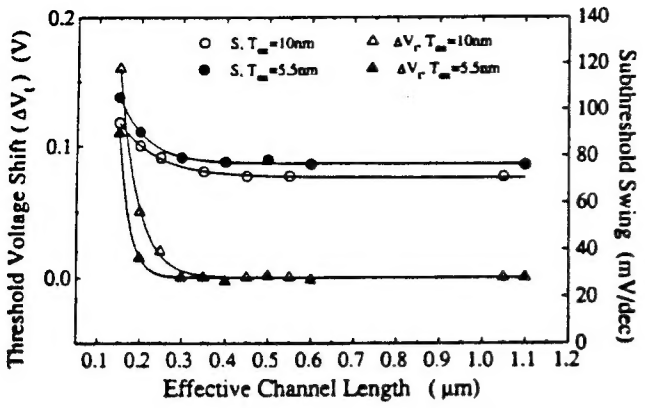


Fig. 2. Threshold voltage shift (ΔV_t) and subthreshold swing (S) versus effective channel length at V_{ds} = −0.1 V. ΔV_t is the difference between V_t of a long-channel device and V_t of the given device.

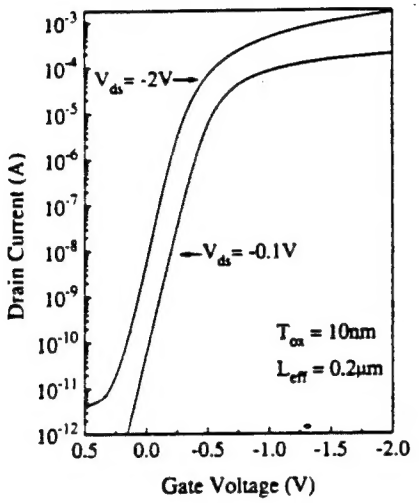


Fig. 3. Subthreshold characteristics of a device with W/L_{eff} = 9.5 μm/0.2 μm.

for series resistance effect. In fact a key to achieved transconductance is the relatively low series resistance, which ranges from 700 to 1100 $\Omega \cdot \mu$m for our devices. The G_m of our devices with T_{ox} = 10 nm is higher than those reported for bulk devices with T_{ox} = 6 nm and T_{ox} = 8 nm [4], [5]. Fig. 5 similarly shows that the I_{dsat} of present devices is larger than recently reported values for both bulk and SOI devices with thinner gate oxides [1], [5], [6].

There are several reasons for SOI devices built on thin silicon films to have higher G_m and I_{dsat} over their bulk counterparts. Fully depleted (FD) and nearly fully depleted (NFD) devices have reduced or no bulk charge effect that raises the local V_t increasingly toward the drain. Reduced bulk charge also reduces the local effective vertical field and improves the carrier mobility [7]. An additional effect not reported before is the effect of forward bias on the floating body *even before the onset of the kink*. To demonstrate this effect, special four-terminal devices with body contacts were fabricated on the same die as regular three-terminal devices. Fig. 6 shows the I–V

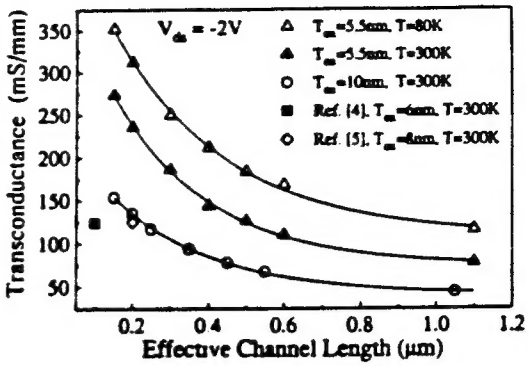


Fig. 4. Measured saturation transconductance (G_m) versus effective channel length. V_{ds} = −2 V.

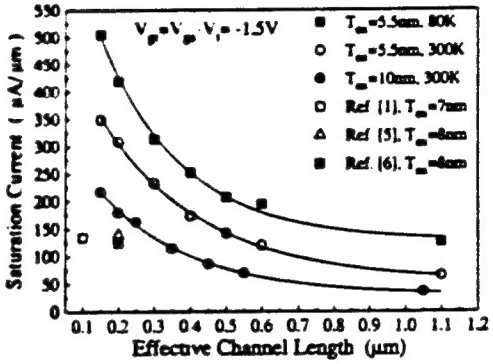


Fig. 5. Normalized saturation current versus effective channel length. V_{gs} = V_{gs} − V_t = −1.5 V.

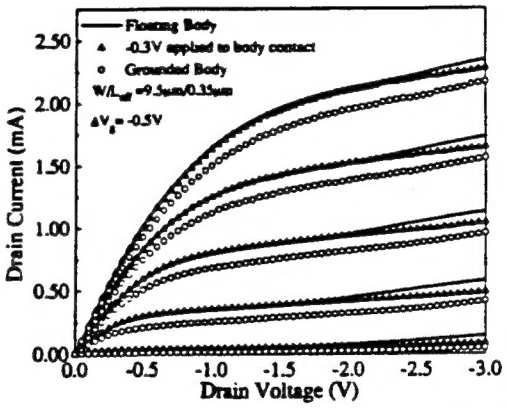


Fig. 6. I–V characteristics of a four-terminal device that has body contact. Solid lines show I–V when body is floating. Circles are used for the grounded body, and solid triangles are used when −0.3 V is applied to body contact. Each V_g step is −0.5 V.

of a four-terminal device. Three sets of curves are shown with body floating, body grounded, and −0.3 V (forward bias) applied to the body. I_d is clearly larger with the fourth terminal open than grounded. However, the floating-body case and the forward-bias case match before onset of the kink, indicating that a forward bias of about 0.3 V is present when the body is floating. Since the body–drain junction is reverse biased and some leakage current flows from drain to body, the forward bias of the

body-source junction allows this current to flow from body to source. One drawback of the forward bias is reduction of threshold voltage and increase of leakage current at $V_g = 0$, as seen in Fig. 3.

IV. Conclusion

Using SIMOX wafers with silicon film thickness of 130 nm, PMOS transistors with effective channel lengths down to 0.15 μm are fabricated. These devices exhibit excellent short-channel behavior, low series resistance, and remarkable G_m and I_{dsat}. For $T_{ox} = 5.5$ nm, G_m of 274 mS/mm at 300 K and 352 mS/mm at 80 K are achieved, which are the highest reported values for this oxide thickness. The high performance is attributed to low series resistance, reduction of body charge effect, and the forward-bias body effect.

References

[1] Y. Omura, S. Nakashima, K. Izumi, and T. Ishii, "0.1-μm-gate, ultrathin-film CMOS devices using SIMOX substrate with 80-nm-thick buried oxide layer," in *IEDM Tech. Dig.*, 1991, pp. 675-678.

[2] S. Parke *et al.*, "A versatile, SOI BiCMOS technology with complementary lateral BJT's," in *IEDM Tech. Dig.*, 1992.

[3] J. Chung *et al.*, "Deep-submicrometer MOS device fabrication using a photoresist-ashing technique," *IEEE Electron Device Lett.*, vol. 9, pp. 186-188, 1988.

[4] M. Aoki *et al.*, "Design and performance of 0.1-μm CMOS devices using low-impurity-channel transistors (LICT's)," *IEEE Electron Device Lett.*, vol. 13, pp. 50-52, 1992.

[5] R. Chapman *et al.*, "High performance sub-half micron CMOS using rapid thermal processing," in *IEDM Tech. Dig.*, 1991, pp. 101-104.

[6] A. Hori, S. Kameyama, M. Segawa, H. Shimomura, and H. Ogawa, "A self-aligned pocket implantation (SPI) technology for 0.2 μm-dual gate CMOS," in *IEDM Tech. Dig.*, 1991, pp. 641-644.

[7] J. C. Sturm, K. Tokunaga, and J. Colinge, "Increased drain saturation current in ultra-thin silicon-on-insulator (SOI) MOS transistors," *IEEE Electron Device Lett.*, vol. 9, pp. 460-463, 1988.

Observation of Velocity Overshoot in Silicon Inversion Layers

Fariborz Assaderaghi, Ping Keung Ko, *Member, IEEE*, and Chenming Hu, *Fellow, IEEE*

Abstract—**Employing a novel test structure, electron velocity overshoot in silicon inversion layers is observed at room temperature. For channel lengths longer than 0.3 μm, the velocity / field relation follows the well-known behavior with no channel length dependence. The first indication of velocity overshoot is seen at channel length of 0.22 μm, while at $L = 0.12$ μm drift velocities up to 35% larger than the long channel value are measured.**

I. Introduction

As MOS transistor dimensions shrink to deep-submicrometer regime, nonlocal effects are expected to become more prominent. Perhaps the most important of these nonlocal effects is velocity overshoot, which can improve current drive and transconductance. Several authors have provided theoretical models (see [1], [2] and references therein). Recently, measurement of very high transconductance in 0.1-μm MOSFET's was attributed to velocity overshoot [3]. This attribution was made by comparing the measured transconductance with Monte Carlo simulation of the reported device [4]. Here we report observation of velocity overshoot, using a special test structure.

II. Device Structure

As in our previous work of measuring saturation velocity [5], we employ back-channel conduction in silicon-on-insulator (SOI) MOSFET's. The SOI devices used in the study are built on SIMOX wafers. A full description of process integration is given in [6]. Here we provide only the key device parameters. As shown in Fig. 1, front-gate oxide thickness T_{fg}, silicon film thickness T_{si}, and buried oxide thickness T_{bg} are 18, 130, and 400 nm, respectively. The doping concentration is approximately $6-8 \times 10^{16}$ cm^{-3}. In the normal mode of operation of these devices, the inversion layer is formed at the front Si/SiO_2 interface. However, it is possible to form the inversion layer at the *buried oxide / silicon* interface by applying a very large back-gate voltage V_{bg}. To eliminate conduction by the front channel, negative front-gate voltage is applied to accumulate the front Si/SiO_2 interface. This unusual structure and bias condition provide a unique opportunity for observing velocity overshoot as follows. For short-channel devices (e.g., 0.5 μm), only a small drain voltage (e.g., 1 V) is required to achieve a high tangential field. Since the drain voltage V_d is much smaller than back-gate voltage V_{bg} (e.g., 70 V), the inversion charge density is essentially uniform in the channel between source and drain. Thus the tangential field is uniform. This is to be contrasted with a regular thin-oxide MOSFET where the tangential field is nonuniform and increases significantly from source to drain.

The idea of utilizing very thick gate oxides to obtain uniform inversion layers was first tried in bulk MOSFET's by Fang and Fowler [7]. They used this technique to measure electron saturation velocity in inversion layers with good accuracy. However, if one employs a submicrometer bulk MOSFET with very thick gate oxide, the device will suffer from punchthrough. In the SOI MOSFET punchthrough is effectively suppressed due to the presence of thin silicon film.

III. Results and Discussion

NMOSFET's with $W = 9.5$ μm and channel lengths from 0.6 to 0.12 μm were used. Front-gate voltage was set to −4 V to accumulate the front interface, while the back-gate threshold voltage was about 11 V. Since V_{bg} was in the range of 60–100 V and V_d was kept below 1.5 V, the inversion charge was essentially uniform, allowing us to write $I = C_{bg}W(V_{bg} - V_t)v$, where C_{bg} is the buried oxide capacitance, V_t is the back gate threshold voltage, W is the channel width, and v is the electron drift velocity. Since v is the only unknown in this relation, it can be determined from the measured current. The tangential field is given by $E_y = (V_d - IR_{sd})/L$, where R_{sd} (series resistance) is 50–60 Ω for our devices.

Fig. 2 shows electron drift velocity versus tangential field for a 0.47-μm device. Very good agreement with Thornber's equation [8] is achieved for the usual choice of $\beta = 2$:

$$V(E_y) = \mu_o E_y \left(1 + \left(\frac{\mu_o E_y}{v_{sat}}\right)^{\beta}\right)^{-1/\beta}.$$

As seen in this figure, the low field mobility at $V_{bg} = 50$ V is 480 $cm^2/V \cdot s$ and decreases to 390 $cm^2/V \cdot s$ at $V_{bg} =$ 90 V, as expected. Not surprisingly, velocity tends to saturate at tangential fields above 3×10^4 V/cm, and it does not show strong dependence on the vertical field.

Manuscript received April 20, 1993; revised August 3, 1993. This work was supported by SRC under Contract 93-DC-324, ISTO/SDIO through ONR under Contract N00014-92-J-1757, and AFOSR/JSEP under Contract F49620-93-C0041.

The authors are with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720.

IEEE Log Number 9212549.

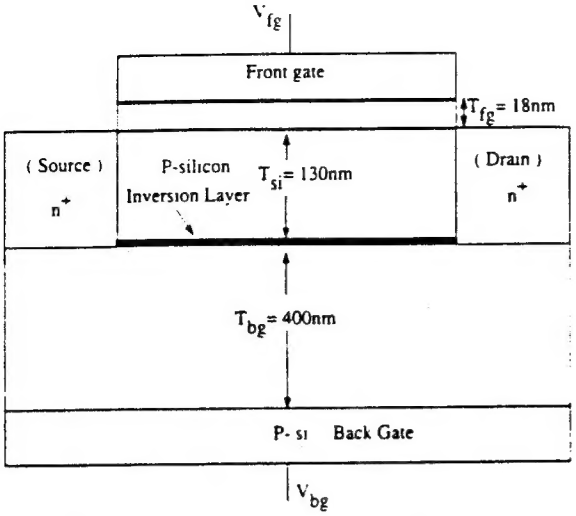


Fig. 1. Schematic cross section of an SOI MOSFET. By applying a large positive voltage to the back gate, the inversion layer is formed at the back Si/SiO_2 interface.

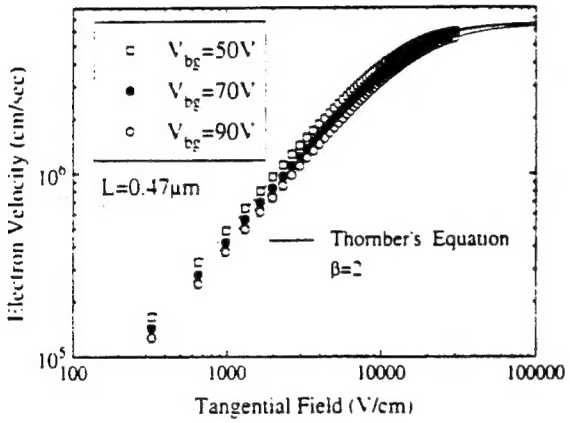


Fig. 2. Measured electron drift velocity versus tangential field for a device with $L = 0.47\ \mu m$. Vertical field is used as a parameter.

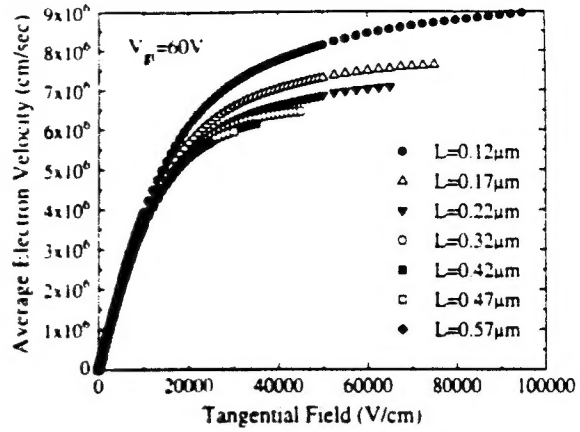


Fig. 3. Measured electron drift velocity versus applied tangential field. Device channel length is used as a parameter. $V_{gt} = V_{bg} - V_t = 60$ V.

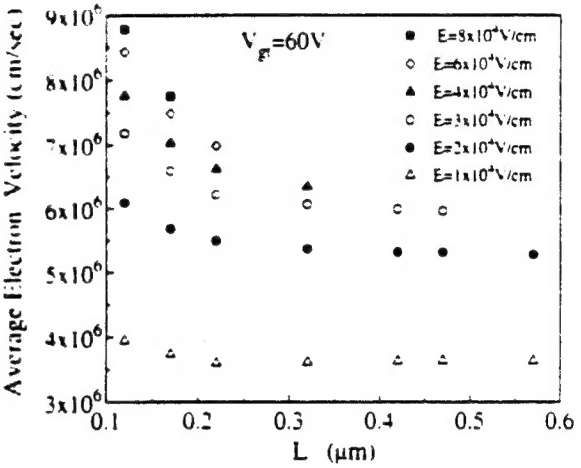


Fig. 4. Average electron drift velocity versus channel length, with tangential field used as a parameter.

Fig. 3 shows the results of similar measurements on devices with different channel lengths. $v(E_y)$ for channel lengths in the range of 0.6–0.35 μm nearly overlap, but for shorter channel lengths the high field velocity starts to increase with decreasing channel length. Clearly for $L < 0.25\ \mu$m, the drift velocity exceeds the saturation velocity of long-channel structures. For example, at $L = 0.12\ \mu$m drift velocities up to 35% larger than the saturation velocity are observed. It should be noted that the concept of uniform charge, electric field, and drift velocity that we used to derive the velocity/field relationship is only valid for long-channel devices (i.e., $L > 0.32\ \mu$m). For very short devices (where overshoot is not negligible), the velocity is not constant in the channel, and the values reported here should be treated as "average" drift velocities.

Fig. 4 shows the average drift velocity as a function of channel length, with the tangential field as a parameter. For $L > 0.25\ \mu$m (e.g., $L = 0.32\ \mu$m), as tangential field increases the drift velocity increases but tends to saturate for larger fields. This is to be contrasted with the $L = 0.12$-μm device, which shows no clear velocity saturation even at $E_y = 8 \times 10^4$ V/cm. Moreover, for relatively moderate fields (e.g., 1×10^4 V/cm), the measured velocities for all different channel lengths are about the same and no significant overshoot is observed. This is obviously not the case for larger fields.

One complicating factor in the above measurements is that for very short channel lengths, the threshold voltage becomes dependent on the drain voltage. Fig. 5 shows I_d–V_g characteristics of the 0.12-μm device, which represents the worst case of V_t reduction. We took into account the V_d dependence of threshold voltage, by measuring current shifts at different drain voltages.

IV. Conclusion

Novel SOI structures are utilized to study the phenomenon of velocity overshoot. Velocity overshoot is observed at room temperature for channel lengths as long as 0.22 μm. At 0.12 μm, drift velocities up to 35% larger than the saturation velocity are measured.

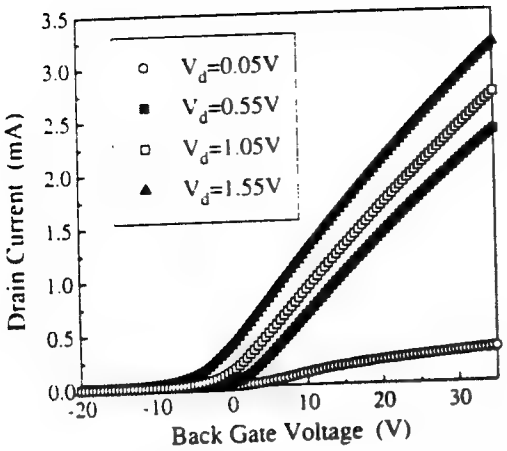


Fig. 5. Drain current of the L = 0.12-μm device plotted as a function of back-gate voltage. Drain voltage is used as a parameter

ACKNOWLEDGMENT

The authors would like to thank the Berkeley Microfabrication Lab. staff in helping with fabrication of the tested structures.

REFERENCES

[1] M. Fischetti and S. Laux, "Monte Carlo analysis of electron transport in small semiconductor devices including band-structure and space-charge effects," *Phys. Rev. B*, vol. 38, pp. 9721–9745, 1988.

[2] Baccarani and M. Wordeman, "An investigation of steady-state velocity overshoot in silicon," *Solid-State Electron.*, vol. 28, pp. 407–416, 1985.

[3] G. Sai-Halasz, M. Wordeman, D. Kern, S. Rishton, and E. Ganin, "High transconductance and velocity overshoot in NOMS devices at the 0.1-μm gate-length level," *IEEE Electron Device Lett.*, vol. 9, pp. 464–466, 1988.

[4] S. Laux and M. Fischetti, "Monte Carlo simulation of submicrometer Si n-MOSFETs at 77 and 300K," *IEEE Electron Device Lett.*, vol. 9, pp. 467–469, 1988.

[5] F. Assaderaghi, J. Chen, P. Ko, and C. Hu, "Measurement of electron and hole saturation velocities in silicon inversion layers using SOI MOSFETs," in *IEEE Int. SOI Conf. Proc.*, 1992, pp. 112–113.

[6] S. Parke, F. Assaderaghi, J. Chen, J. King, C. Hu, and P. Ko, "A versatile, SOI BiCMOS technology with complementary lateral BJT's," in *IEDM Tech. Dig.*, 1992.

[7] F. F. Fang and A. B. Fowler, "Hot electron effects and saturation velocities in silicon inversion layers," *J. Appl. Phys.*, vol. 41, pp. 1825–1831, 1970.

[8] K. K. Thornber, "Relation of drift velocity to low-field mobility and high-field saturation velocity," *J. Appl. Phys.*, vol. 51, pp. 2127–2136, 1980.

93CH3330-8

1993 IEEE INTERNATIONAL SOI CONFERENCE PROCEEDINGS

October 5-7, 1993
The Autry Resort
Palm Springs
California

General Chairman: John Schott
Technical Program Chairman: Witek P.Maszara
Local Arrangements Chairman: Jay Schrankler
Tresurer/Registration Chairman: Donald C.Mayer
Rump/Poster Session Chairman: Scott M.Tyson

ROOM TEMPERATURE OBSERVATION OF VELOCITY OVERSHOOT IN SILICON INVERSION LAYERS

Fariborz Assaderaghi, Ping Keung Ko, and Chenming Hu
Department of Electrical Engineering and Computer Sciences
U. C. Berkeley, Berkeley CA 94720

As MOS transistor dimensions shrink to deep sub-micron regime, the non-local effects are expected to become more prominent. Perhaps the most important of these non-local effects is velocity overshoot, which can be beneficial to device performance by improving current drive and transconductance. Here, for the first time, we report direct observation of velocity overshoot using a special test structure. The first indication of velocity overshoot is seen at channel length of 0.22μm, while at L_{eff}=0.12μm drift velocity values up to 40% higher than the long channel value are measured.

As in our previous work of measuring saturation velocity [1], we employ back channel conduction in silicon-on-insulator (SOI) MOSFET's. The SOI devices used in the study are built on SIMOX wafers [2]. As shown in Fig.1, the front gate oxide thickness (T_{fg}), silicon film thickness (T_{si}), and buried oxide thickness (T_{bg}) are 18nm, 130nm, and 400nm, respectively. The film doping concentration is approximately 6-8x10^{16} cm^{-3}. The inversion layer is formed at the *buried oxide/silicon* interface by applying a very large back gate voltage V_{bg}. To eliminate conduction by the front channel, negative front gate voltage is applied to accumulate the front Si/SiO_2 interface. This unusual bias condition provides a unique opportunity for observing velocity overshoot as follows.

NMOSFET's with channel lengths from 0.6μm to 0.12μm were used. Since back-gate voltage V_{bg} was in the range of 60-100V and V_d was kept below 1.5V, the inversion charge was essentially uniform between source and drain [3], allowing us to write: $I=C_{bg}W(V_{bg}-V_t)v$. Where C_{bg} is the buried oxide capacitance, V_t is the back gate threshold voltage, W is the channel width, and v is the electron drift velocity. Since v is the only unknown in this relation, it can be determined from the measured current. The tangential field is simply: $E_y=(V_d-IR_{sd})/L_{eff}$. For the measured devices W=9.5μm and R_{sd}=55Ω, making the correction term IR_{sd} small, around 0.1V. Using above relations, in Fig.2 drift velocity is plotted versus tangential field for a 0.47μm device. This velocity/field data is in perfect agreement with the well known Thornber's equation [4]. Moreover, velocity tends to saturate at tangential fields above 3x10^4V/cm, and it does not show strong dependence on the vertical field, as expected.

Fig.3 shows the results of similar measurements on devices with different channel lengths. Drift velocities ($v(E_y)$) for channel lengths longer than 0.32μm nearly overlap. However, for channel lengths shorter than 0.32μm this overlap disappears, and the high-field velocity starts to increase with decreasing channel length. Clearly, for L < 0.25μm, the drift velocity exceeds the saturation velocity of long channel structures. For example, at L=0.12μm drift velocities up to 40% larger than saturation velocity are observed. In fact, at this channel length the drift velocity shows no clear saturation behavior even at E_y = 1x10^5V/cm. Fig.4 shows drift velocity as a function of channel length, with the tangential field as a parameter. For low fields the measured velocities for all different channel lengths are about the same, and no significant overshoot is observed. This is obviously not the case for larger fields.

It should be noted that the idea of uniform charge, field, and velocity that we utilized to derive velocity/field relationship is only valid for long channel devices. For very short channels, the velocity is not constant in the channel (due to velocity overshoot) and the measured value is an average velocity.

0-7803-1346-1/93/$3.00 © 1993 IEEE

Acknowledgment:

This project was supported by SRC under contract number 93-DC-324, ISTO/SDIO through ONR under contract number N00014-92-J-1757, and AFOSR/JSEP under contract number F49620-90-C-0029.

References

[1] F. Assaderaghi et al., *IEEE Int. SOI Conf. Proceedings*, 1992, pp. 112-113.

[2] S. Parke et al., *IEDM Tech. Dig.*, 1992. pp. 453-456.

[3] F. F. Fang and A. B. Fowler, *J. Appl. Phys.*, vol. 41, pp. 1825-1831, 1970.

[4] K. K. Thornber, *J. Appl. Phys.*, vol. 51, pp. 2127-2136, 1980.

[5] G. Baccarani, and M. Wordeman, *Solid-St. Electron.*, vol. 28, pp. 407-416, 1985.

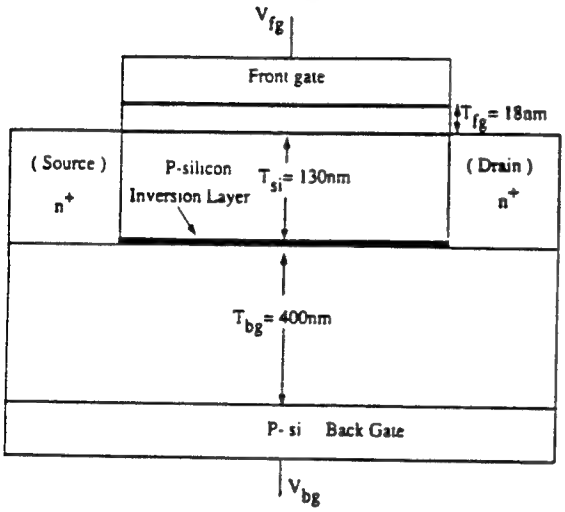


Fig.1 Schematic cross-section of an SOI MOSFET. The inversion layer is formed at the back Si/SiO_2 interface.

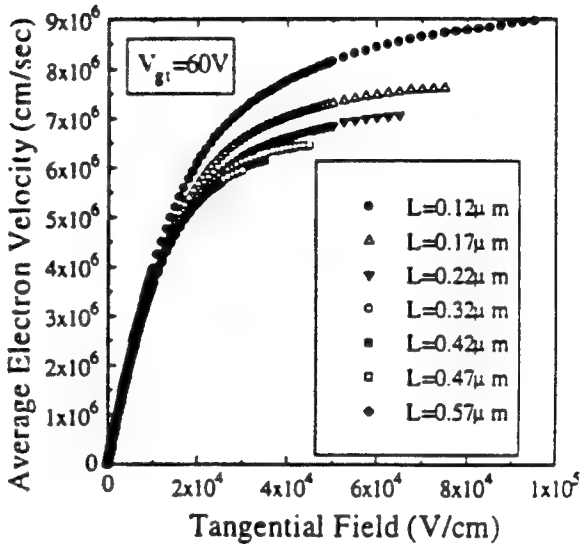


Fig. 3 Average electron velocity versus applied tangential field, with device channel length as a parameter.

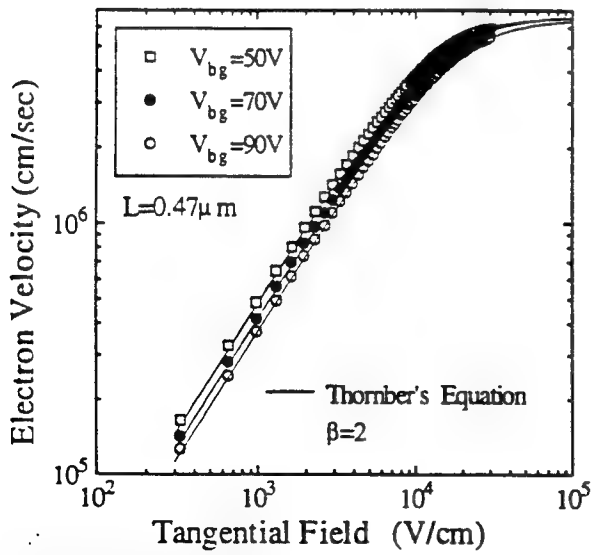


Fig. 2 Measured electron drift velocity versus applied tangential field. Vertical field is used as a parameter.

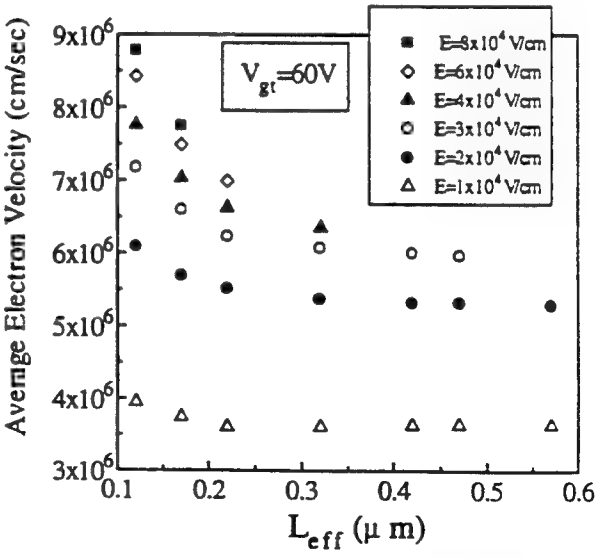


Fig. 4 Average electron drift velocity versus channel length, with tangential field used as a parameter.

STUDY OF CURRENT DRIVE IN DEEP SUB-MICROMETER SOI PMOSFET'S

Fariborz Assaderaghi, Kelvin Hui, Stephen Parke, Jon Duster, Ping K. Ko, and Chenming Hu
Department of Electrical Engineering and Computer Science
University of California, Berkeley CA 94720

ABSTRACT

Sub-quarter micrometer PMOSFET's are fabricated on SOI films, exhibiting excellent short channel behavior, low source-drain resistance, and remarkably large current drive and transconductance. For T_{ox}=5.5nm, saturation transconductances of 270mS/mm at 300K and 350mS/mm at 80K are achieved, which are the highest reported values for this oxide thickness. Direct measurements and simulation results show that the improved current drive is due to low series resistance, forward bias body effect, and the reduction of body charge effect.

INTRODUCTION

Process simplicity and other advantages such as reduced parasitic capacitance and improved short channel effect have led to the development of silicon-on-insulator (SOI) MOSFET's. It has also been shown that long channel SOI MOSFET's have larger current drive than bulk MOSFET's. However, current drive advantage of deep sub-micron SOI MOSFET's over bulk devices has not been demonstrated. Often high series resistance has obscured this advantage [1]. Here, for the first time we report experimental results for deep sub-micron SOI PMOSFET's with improved performance over their bulk counterparts.

SILICON FILM THICKNESS CONSIDERATION

It is difficult to achieve a large enough threshold voltage in a polysilicon gate fully-depleted SOI device in ultra-thin silicon film. In addition, the threshold voltage is very sensitive to film thickness variation. Also, silicidation is necessary to reduce the series resistance. Nearly-Fully-Depleted (NFD) devices can provide the SOI advantages without the above drawbacks.

Based on this consideration, SIMOX substrates with SOI film thickness of 130nm were used. Mesas were created by plasma etching a nitride/oxide/silicon stack. Next, a 100nm oxide was grown on the mesa sidewalls to prevent low-V_t edge devices and gate oxide defects at the mesa corners. Threshold implants were then performed, resulting in concentration of 1-3x$10^{17}$$cm^{-3}$. Gate oxides of 5.5nm and 10nm thickness were grown, followed by the deposition of 280nm of undoped polysilicon. The poly gate was doped by 5x$10^{15}$$cm^{-2}$ low energy boron implant.

The combination of P^+-poly gate and silicon film thickness and doping concentration resulted in NFD devices with the threshold voltage range of -0.3V to -0.5V. Effective channel lengths as short as 0.08μm were obtained by O_2 "ashing" of the gate photoresist [2].

DEVICE PERFORMANCE

Fig. 1 shows the I-V characteristics of a 9.5μm/0.2μm device with T_{ox}=5.5nm. Although the fabricated devices are NFD, and long channel devices show a small kink in their I-V, very short channel devices have no observable kink as seen in Fig.1. This is due to the fact that the depletion regions of the source/drain junctions effectively deplete the film at drain voltages lower than onset of the kink. This is demonstrated by PISCES simulation of a PMOSFET with L_{eff} =0.2μm. As seen in Fig.2, the barrier potential at the bottom of the silicon film is

lowered, reducing the number of electrons that can accumulate in this potential well.

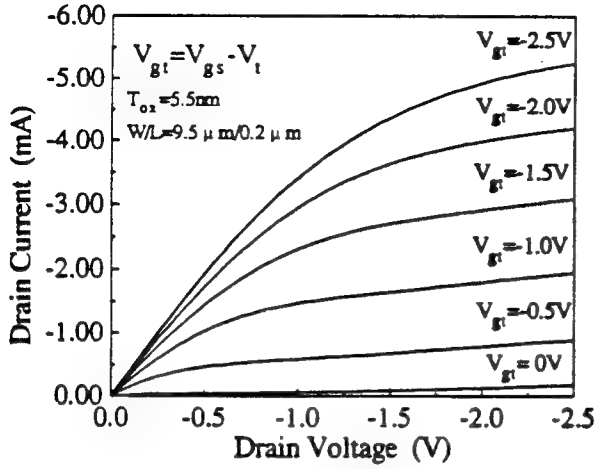


Fig. 1 I-V charactenstics of an NFD PMOSFET. No observable kink is present.

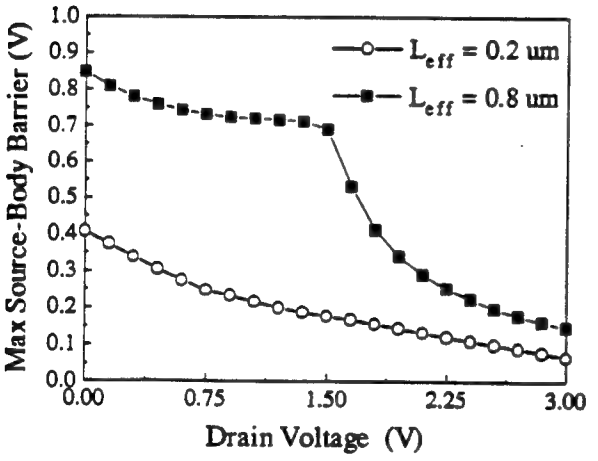


Fig. 2 PISCES simulation of maximum barrier potential at the bottom of silicon film (at buriedoxide interface). Silicon film thickness is 130nm.

Fig.3 shows the subthreshold swing and threshold voltage shift (ΔV_t) of the fabricated PMOSFET's. Although ultra- thin film is not used, good subthreshold swing (75-80mV/dec) and excellent short channel behavior is obtained. Virtually no threshold voltage shift is observed for devices down to 0.2μm. Fig. 4 shows the subthreshold characteristics of a 9.5μm/0.2μm device with T_{ox}=10nm.

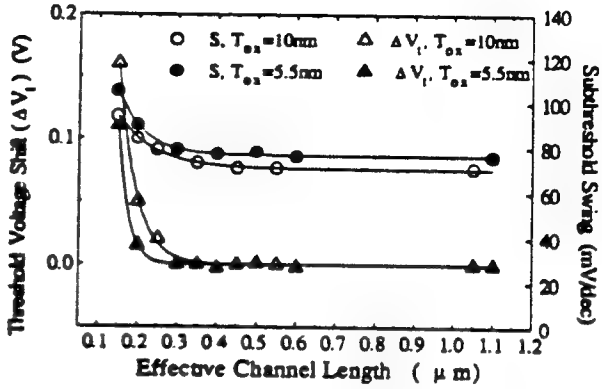


Fig. 3 ΔV_t and S versus effective channel length. ΔV_t is the difference between V_t of a long channel device and V_t of the given device V_{ds}= -0.1V.

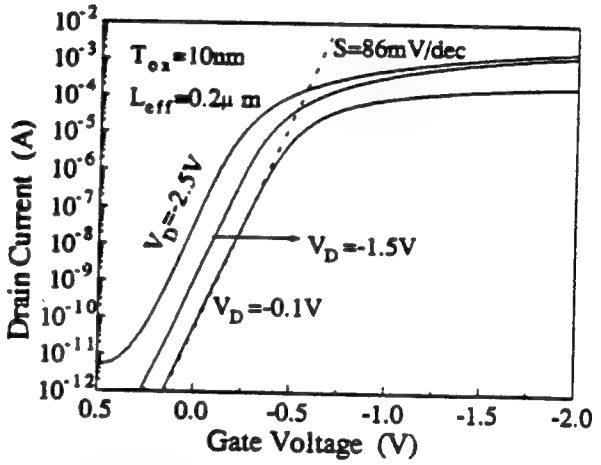


Fig. 4 Subthreshold characteristics of an NFD PMOSFET with L_{eff} =0.2μm.

Fig.5 shows that for T_{ox}= 5.5nm, the device with L_{eff} of 0.15μm has G_m of 274mS/mm at 300K and 352mS/mm at 80K. These are measured values and have not been corrected for series resistance effect. In fact a key to achieved transconductance is the relatively low series resistance, which ranges from 700Ωμm to 1100Ωμm for our devices. The G_m of our devices with T_{ox}=10nm is higher than those reported for bulk devices

with T_{ox}=6nm and T_{ox}=8nm [3,4]. Fig.6 similarly shows that I_{dsat} of present devices is larger than recently reported values for both bulk and SOI devices with thinner gate oxides [1,4,5].

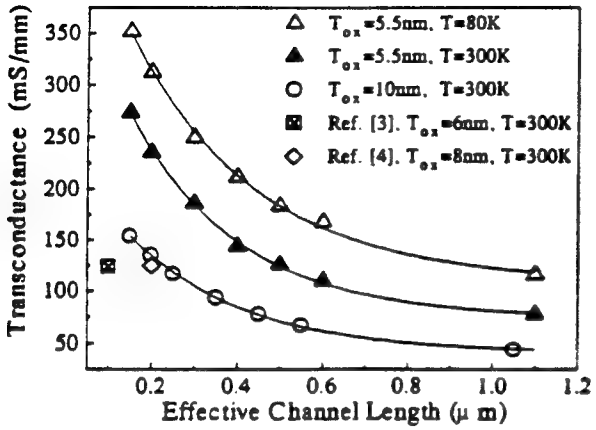


Fig 5 Measured saturation transconductance versus effective channel length.

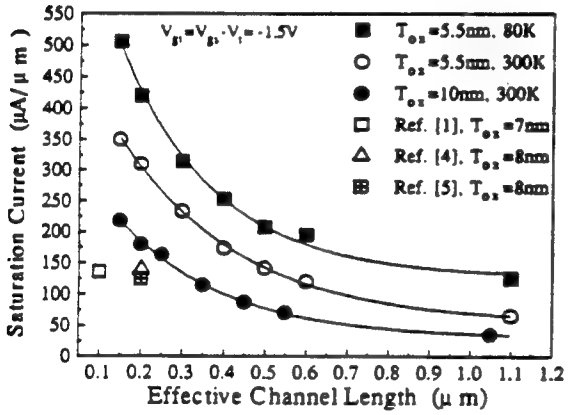


Fig. 6 Measured saturation current (I_{dsat}) versus effective channel length. $V_{gt} = V_{gs} - V_t = -1.5V$.

There are several reasons for SOI devices built on thin silicon films to have higher G_m and I_{dsat} over their bulk counterparts. Fully-depleted (FD) and Nearly-Fully-Depleted (NFD) devices have reduced or no bulk charge effect that raises the local V_t increasingly toward the drain. Reduced bulk charge also reduces the local effective vertical field and improves the carrier mobility [6]. Simulation of this effect is shown for a 0.5μm PMOSFET in Fig. 7. As seen the SOI device has a larger current drive than the bulk MOSFET. The bulk charge effect, however, becomes a smaller factor for shorter devices. Fig.8 shows that the improvement of SOI current drive over bulk due to this component disappears at L_{eff} =0.1μm.

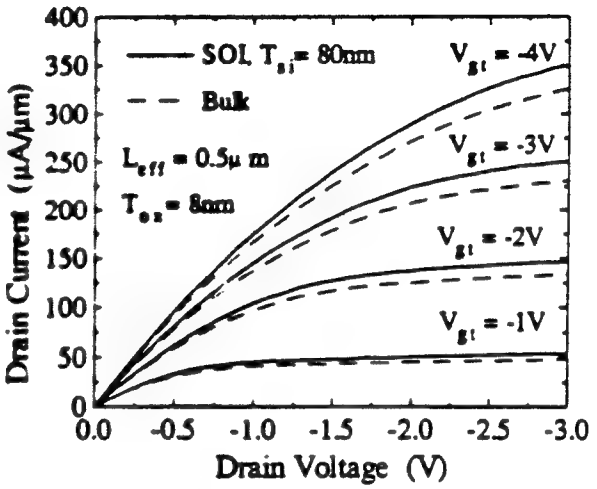


Fig 7 PISCES simulation of body charge effect for a 0.5μm PMOSFET. The SOI device has larger current drive due to absence of bulk charge effect.

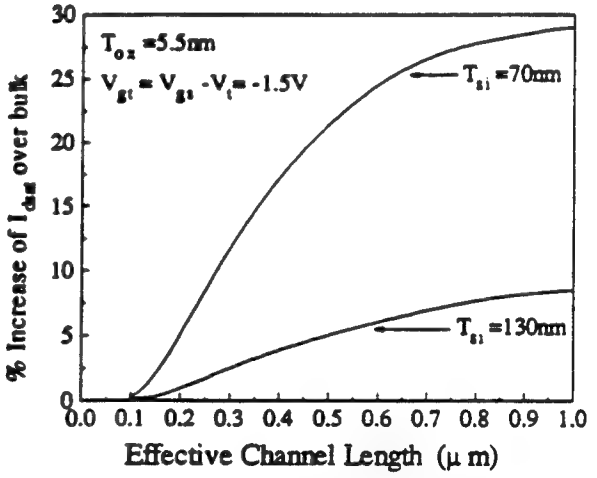


Fig. 8 PISCES simulation of improvement in SOI current drive over bulk due to reduction of body charge efffect. Two film thicknesses are simulated.

An additional effect not reported before is the effect of forward bias on the floating body *even before the onset of the kink*. Fig.9 demonstrates this with a four terminal SOI device that has a body contact. I_d is clearly larger with the fourth terminal (body contact) open than grounded. In Fig.10, I-V curves for the floating body case are compared with the case where -0.3V is applied to the body contact. This two set of curves match before the kink, indicating that a forward bias of about 0.3V is present when the body is floating (and before onset of the kink). At this body voltage the drain to body reverse leakage current is equal to the body to source forward current as shown in Fig 11.

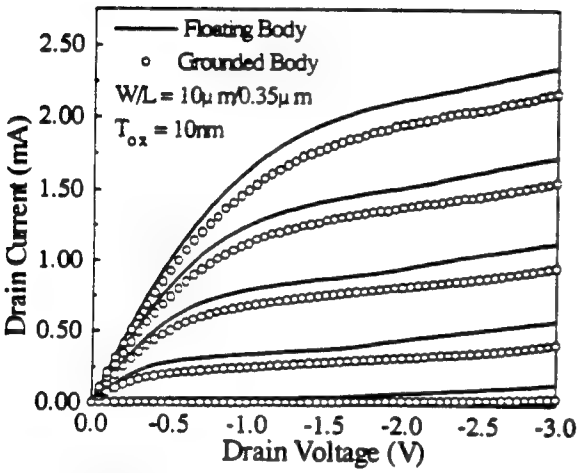


Fig. 9 Circles show I-V when body is grrounded. Solid lines show I-V when body is floating. V_g steps are -0.5V.

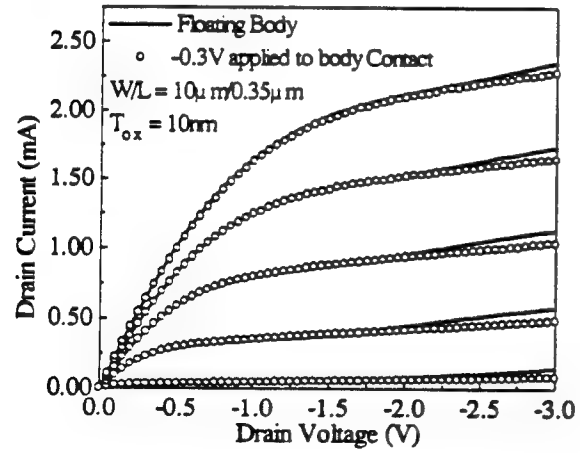


Fig. 10 Circles show I-V when -0.3V is applied to the body. Lines show I-V when body is floating. V_g steps are -0.5V each.

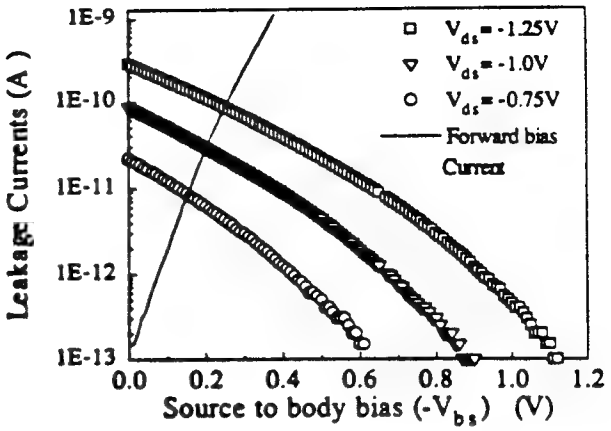


Fig 11 Symbols represent the reverse leakage between drain and body. The line shows the forward current between body and source. Intersects determine the amount of forward bias

The deviation of the two set of curves in Fig. 10 --after the kink-- is due to the amplification of substrate current by the parasitic lateral BJT (for the floating body case). Current gain of the lateral BJT can be significant as seen in Fig. 12, where above PMOSFET is operated as a PNP transistor with the body contact used as the base contact.

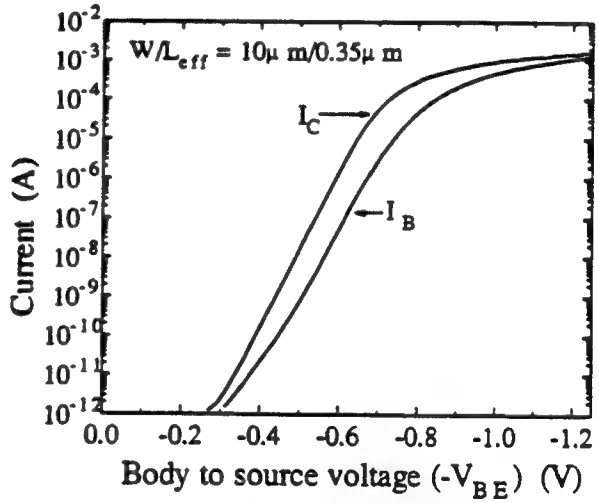


Fig. 12 A four terminal SOI PMOSFET operated as a PNP transistor. Current gain > 30.

SUMMARY

Nearly-Fully-Depleted (NFD) PMOSFET's with effective channel lengths down to 0.15μm were fabricated on 130nm thick SOI films. These devices exhibit excellent short channel behavior, low series resistance, and remarkable G_m and I_{dsat}. For T_{ox}= 5.5nm, G_m of 270mS/mm at 300K and 350mS/mm at 80K are achieved, which are the highest reported values for this oxide thickness. This result is attributed to low series resistance, forward bias body effect, and the reduction of body charge effect.

ACKNOWLEDGEMENT

This work was supported by SRC, IFTO/FDIO through ONR under contract number N00014-85-K-0603 and AFOSR/ JSEP under contract number F49620-87-C-0041. The authors would like to thank the Berkeley Microfabrication Lab. staff for their support in device processing and fabrication.

REFERENCES

[1] Y. Omura, S. Nakashima, K. Izumi, and T. Ishii, "0.1mm-gate, ultrathin-film CMOS device using SIMOX substrate with 80-nm-thick buried oxide layer," IEDM Tech. Dig., 1991, p. 675.

[2] S. Parke, F. Assaderaghi, J. Chen, J. King, C. Hu, and P. Ko, "A versatile SOI BiCMOS technology with complementary lateral BJT's," IEDM Tech. Dig., 1992, p. 453.

[3] M. Aoki, T. Ishii, T. Yoshimura, Y. Kiyota, S. Iijima, T. Yamanaka, T. Kure, K. Ohyu, T. Nishida, S. Okazaki, K. Seki, and K. Shimohigashi, "Design and performance of 0.1-mm CMOS devices using low-impurity-channel transistors (LCT's)," IEEE EDL, vol. 13, p. 50, 1992.

[4] R. Chapman, W. Kuehne, P. Ying, W. Richardson, A. Peterson, A. Lane, I-C. Chen, C. Blanton, M. Moslehi, and J. Paterson, "High performance sub-half micron CMOS using rapid thermal processing," IEDM Tech. Dig., 1991, p. 101.

[5] A. Hori, S. Kameyama, M. Segawa, H. Shimomura, and H. Ogawa, "A self-aligned pocket implantation (SPI) technology for 0.2mm-dual gate CMOS," IEDM Tech. Dig., 1991, p. 641.

[6] J. Sturm, K. Tokunaga, and J. Colinge, "Increased drain saturation current in ultra-thin silicon-on-insulator (SOI) MOS transistors," IEEE EDL, vol. 9, p. 460, 1988.

Variable Threshold Voltage MOSFET (VTMOS) for Very Low Voltage Operation

Fariborz Assaderaghi, Stephen Parke[1], Dennis Sinitsky, Ping. K. Ko, and Chenming Hu

Department of Electrical Engineering and Computer Science
University of California at Berkeley
Berkeley, CA 94704

Abstract

A new mode of operation for Silicon-On-Insulator (SOI) MOSFET is experimentally investigated. This mode gives rise to a Variable Threshold voltage MOSFET (VTMOS). VTMOS threshold voltage drops as gate voltage is raised, resulting in a much higher current drive than regular MOSFET at low V_{dd}. On the other hand, V_t is high at V_{gs}=0, thus the leakage current is low. Suitability of this device for ultra low voltage operation is demonstrated by ring oscillator performance down to V_{dd} = 0.5V.

---

[1]Now with IBM Corporation at East Fishkill, NY

Introduction

During the past few years demand for low power and high performance digital systems has grown rapidly. The main approach for reducing power has relied on power supply scaling. Since power supply reduction below $3V_t$ will degrade circuit speed significantly, scaling of power supply should be accompanied by threshold voltage reduction. However, the lower limit for threshold voltage is set by the amount of off-state leakage current that can be tolerated (due to standby power consideration in static circuits, and avoidance of failure in dynamic circuits and memory arrays). To extend the lower bound of power supply, we propose a Variable Threshold voltage MOSFET (VTMOS) with the highest V_t at zero bias and the lowest value at $V_{gs}=V_{dd}$. In the remainder of this paper we will describe the operation of the device, and show its superiority over a regular MOSFET. We will also show some circuit performances using VTMOS.

Experiment and Results

The SOI devices used in the study are built on SIMOX wafers. Mesa active islands (MESA) were created by plasma-etching a nitride/oxide/silicon stack stopping at buried oxide. P+ polysilicon gate was used for PMOSFET's and N+ for NMOSFET's. A four terminal layout was used to provide separate source, drain, gate, and body contacts. In addition to the four-terminal layout, devices with local gate-to-body connections were also fabricated as illustrated in Fig. 1. This connection uses an oversized metal to P+ contact window aligned over a "hole" in the poly gate [1]. The metal shorts the gate and P+ region. Thus, there is no significant penalty in area.

To operate the VTMOS, floating body and gate of a Silicon-On-Insulator (SOI) MOSFET are tied together. This is not a new configuration, as [1-3] have already suggested it. However, [1-3] all tried to exploit the extra current produced by the lateral bipolar transistor. This normally requires the body voltage to be larger than 0.6V. Since current gain of the bipolar device is small, extra drain (collector) current comes at cost of excessive input (base) current, which contributes to

the standby current. We will show that most of the improvement can be achieved when gate and body voltages are kept below 0.6V. This also ensures that base current will stay negligible. Although the same idea can be used in bulk devices, better advantage is reached in SOI, where because of very small junction areas base current and capacitances are appreciably reduced.

Fig. 2 illustrates the NMOS behavior, with a separate terminal used to control the body voltage. The threshold voltage at zero body bias is denoted by V_{to}. Body bias effect is normally studied in the reverse bias regime, where threshold voltage increases as body to source reverse bias is made larger. We propose to use the exact opposite regime. Namely, we "forward bias" the body-source junction (at less than 0.6V), forcing the threshold voltage to drop.

Specifically, this forward bias effect is achieved by connecting the gate to the body. This is shown as $V_{gs}=V_{bs}$ line in Fig. 2. The intersect of V_t curve and $V_{gs}=V_{bs}$ line determines the point where gate and threshold voltages become identical. This point, which is marked as V_{tf}, is the VTMOS threshold voltage. This lower threshold voltage does not come at expense of higher off-state leakage current, because at $V_{bs}=V_{gs}=0$ VTMOS and regular device have the same V_t. In fact, they are identical in all respects and consequently have the same leakage. This is clearly seen in Fig. 3. Reduced V_{tf} compared to V_{to} is attained through a theoretically ideal subthreshold swing of 60mV/dec. Fig. 3 demonstrates this for PMOS and NMOS devices operated in VTMOS mode and in regular mode. Subthreshold swing is 80mV/dec in the regular devices.

This is not the only improvement. As the gate of VTMOS is raised above V_{tf}, threshold voltage drops further. The threshold voltage reduction continues until $V_{gs}=V_{bs}$ reaches $2\Phi_b$, and threshold voltage reaches its minimum value of $V_{t,min}=2\Phi_b+V_{fb}$. For example, for technology-B in Fig. 2, at $V_{gs}=V_{bs}=0.6V$, $V_t=0.18V$ compared to $V_{to}=0.4V$. In VTMOS operation the upper bound for applied $V_{gs}=V_{bs}$ is set by the amount of base current that can be tolerated. This is illustrated in Fig. 3, where PMOS and NMOS device body (base) currents are shown. At $V_{gs}=0.6V$ base currents for both PMOS and NMOS devices are less than 2nA/μm. A further advantage of VTMOS is that its carrier mobility is expected to be higher because the depletion charge is reduced and the effective normal field in the channel is lowered [4].

Current drives of VTMOS and regular MOSFET are compared in Fig. 4, for technology-B of Fig. 2. VTMOS drain current is 2.5 times of regular device at V_{gs}=0.6V, and 5.5 times of regular device at V_{gs}=0.3V. AC performance of VTMOS is evaluated by an unloaded 101 stage CMOS ring oscillator. Fig. 5 plots the delay of each stage versus power supply. We emphasize that since the threshold voltages of devices used in the ring oscillator were high (technology-A), the optimum performance was not achieved. For technology-B, ring oscillators are not available. If the devices based on technology-B are used, the expected delay for unloaded ring oscillator can be calculated by the following equation [5]: $\tau_{pd} = \frac{C}{4} V_{dd} (\frac{1}{I_{dsatn}} + \frac{1}{I_{dsatp}})$. This is shown as the dashed line in Fig. 5, where C=200fF is used for W_n=5μm and W_p=10μm. This value for C was obtained by fitting the equation to the measured τ_{pd} of technology-A.

Conclusion

For low power operation at very low voltage, a MOSFET should ideally have a high V_t at V_{gs}=0 to achieve low leakage and low V_t at V_{gs}=V_{dd} to achieve high speed. By tying body and gate of an SOI MOSFET together, a variable threshold voltage MOSFET (VTMOS) is obtained. This device has ideal 60mV/dec subthreshold swing. VTMOS threshold voltage drops as gate voltage is raised, resulting in much higher current drive than regular MOSFET. VTMOS is ideal for very low voltage (< 0.6V) operation, as demonstrated by ring oscillator data. VTMOS also solves the floating body problems of SOI MOSFET such as kinks and V_t stability. Furthermore, carrier mobility is enhanced.

Acknowledgment

This project was supported by SRC under Contract 93-DC-324, ISTO/SDIO through ONR under Contract N00014-92-J-1757, and AFOSR/JSEP under Contract F49620-93-C0041.

References

[1] S. A. Parke, C. Hu, and P. K. Ko, "Bipolar-FET hybrid-mode operation of quarter-micrometer SOI MOSFET's," *IEEE Electron Device Lett.*, vol. 14, no. 5, pp. 236-238, May 1993.

[2] J. P. Colinge, "An SOI voltage-controlled bipolar-MOS device," *IEEE Trans. Electron Devices*, vol. ED-34, no. 4, pp. 845-849, Apr. 1987.

[3] S. Verdonckt-Vandebroek, S. Wong, J. Woo, and P. Ko, "High-gain lateral bipolar action in a MOSFET structure," *IEEE Trans. Electron Devices*, vol. 38, no. 11, pp. 2487-2496, Nov. 1991.

[4] A. G. Sabnis and J. T. Clemens, "Characterization of the electron mobility in the inverted <100> Si surface," in *Int. Electron Devices Meet. Tech. Dig.*, pp. 18-21, Dec. 1979.

[5] C. Hu, "Low-voltage CMOS device scaling," in *IEEE Int. Solid-State Circuit Conf. (ISSCC) Digest of Technical Papers*, pp. 86-87, Feb. 1994.

Figure Captions

Fig. 1 a) Cross section of an SOI NMOSFET with body and gate tied together. b) Gate to body connection by using aluminum to short the gate and P+ region.

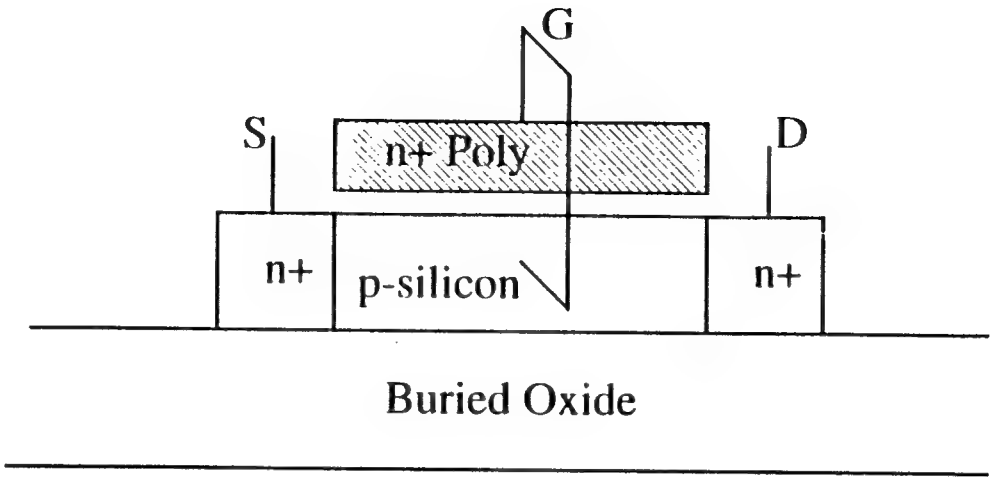
Fig. 2 Threshold Voltage of SOI NMOSFET as a function of body-source forward bias. For Technology-A T_{ox}=10nm, N_a=2.0x10^{17}cm^{-3}. For Technology-B T_{ox}=6.4nm N_a=2.3x10^{17}cm^{-3}.

Fig. 3 Subthreshold characteristics of SOI NMOSFET and PMOSFET operated with body grounded and body tied to the gate. Body to source currents are also shown for the case of VTMOS (body tied to the gate).

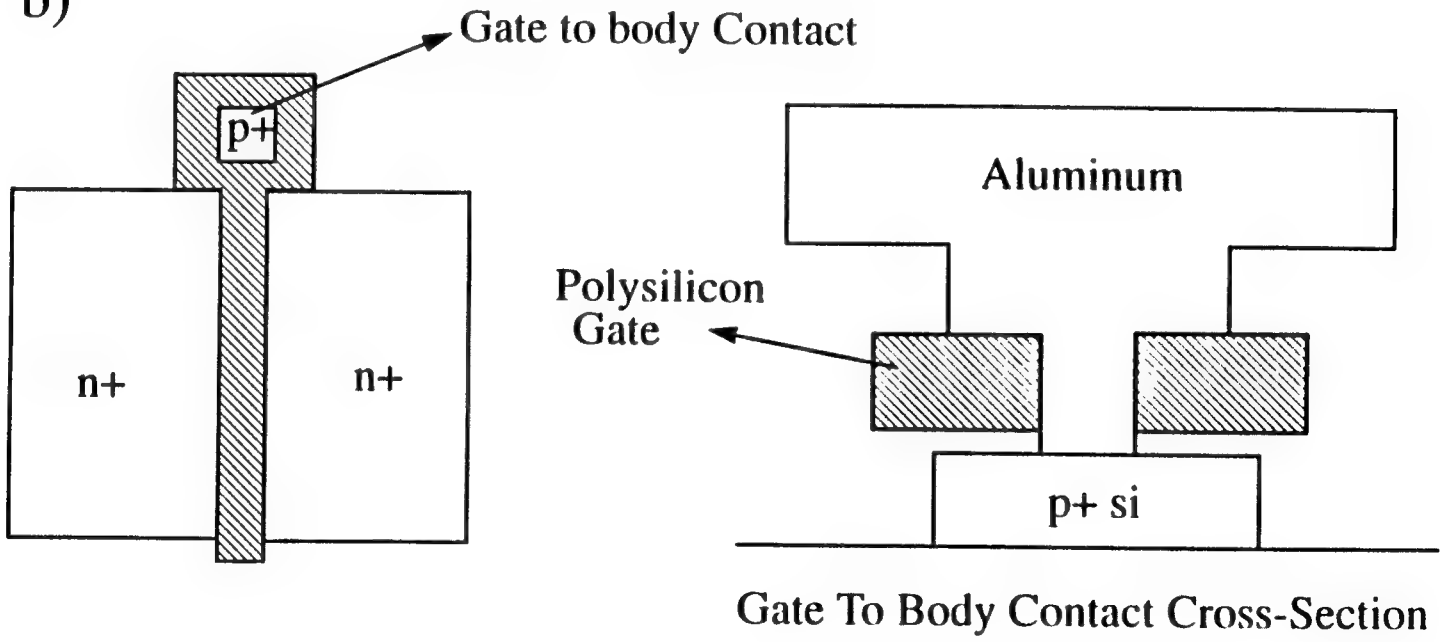
Fig. 4 Drain current of an SOI NMOSFET operated as a VTMOS and as a regular device.

Fig. 5 Delay of a 101-stage ring oscillator. The PMOS and NMOS devices in the ring are VTMOS with T_{ox}=10nm, and L_{eff}=0.3μm, V_{to}=0.6V. The dashed line is prediction of delay for a ring oscillator based on Technology-B with L_{eff}=0.3μm.

a)



b)



Gate To Body Contact Cross-Section

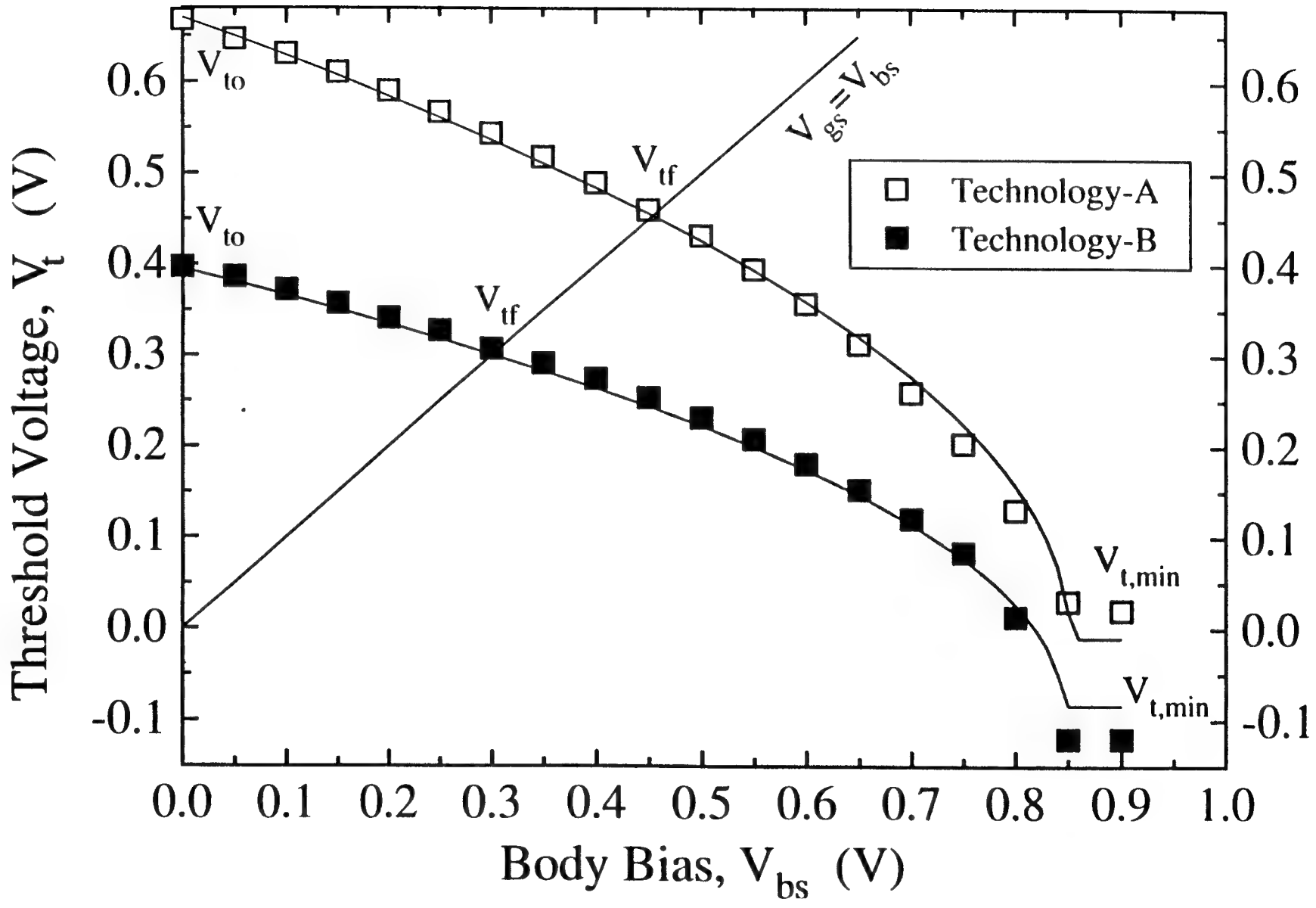
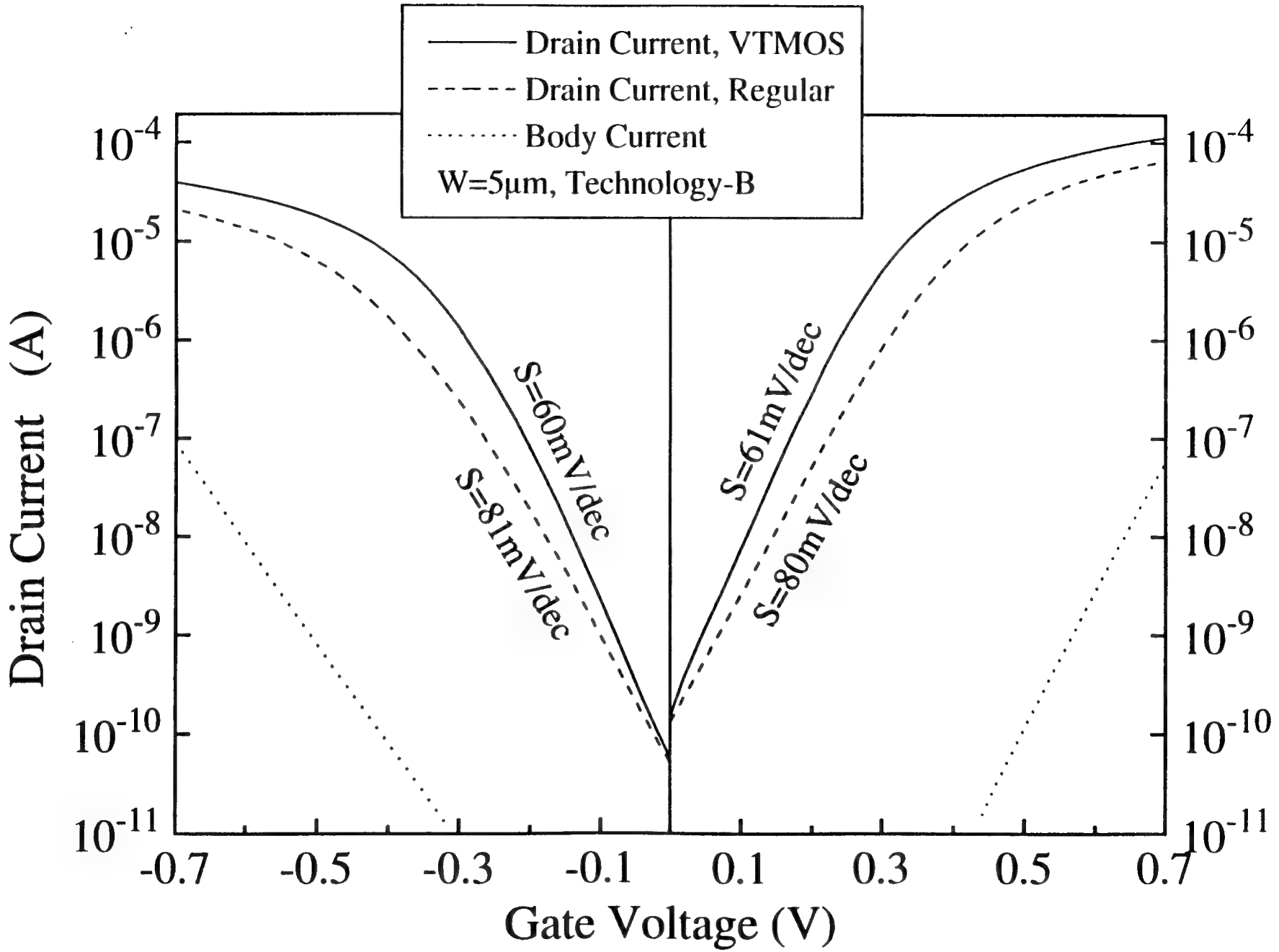
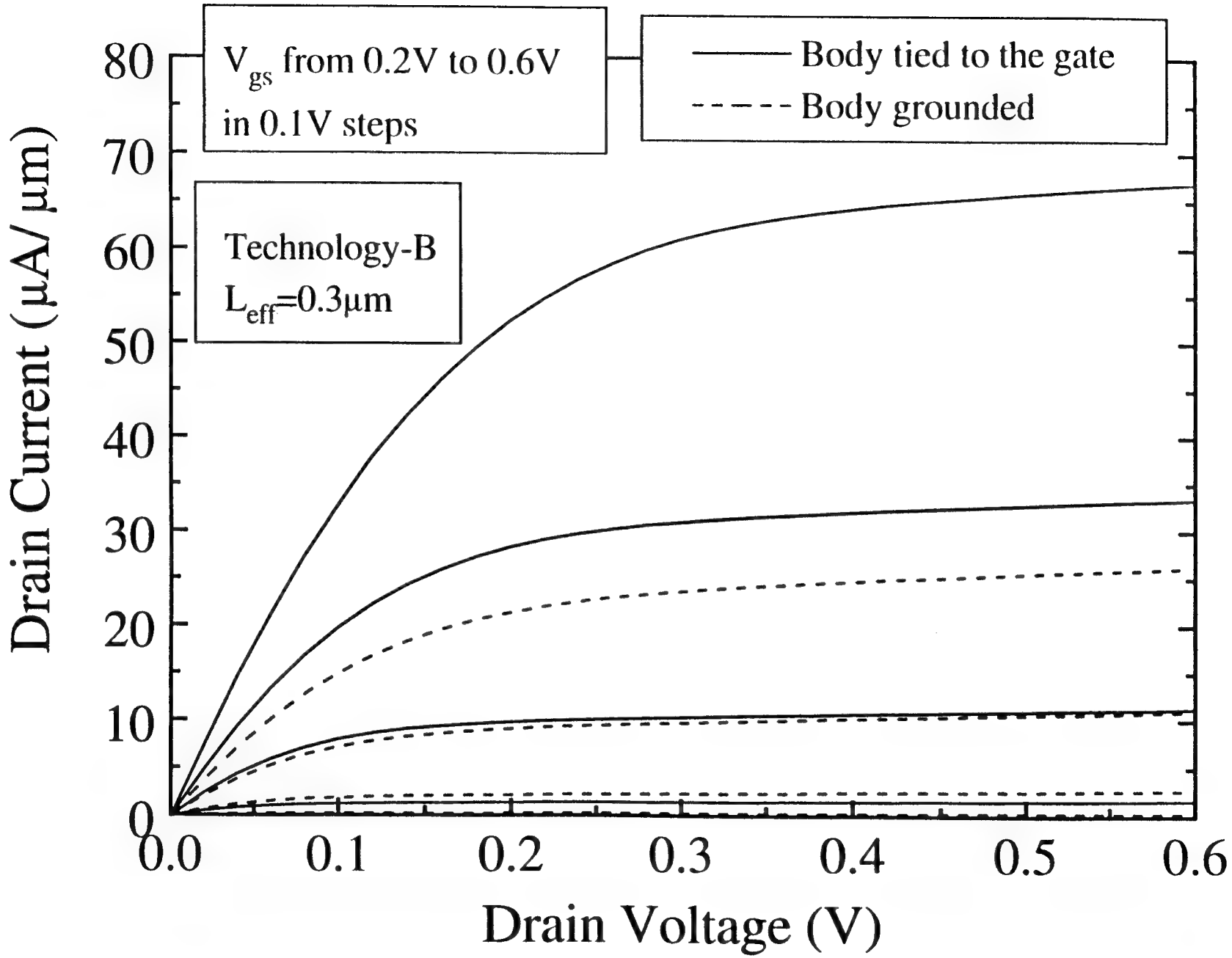
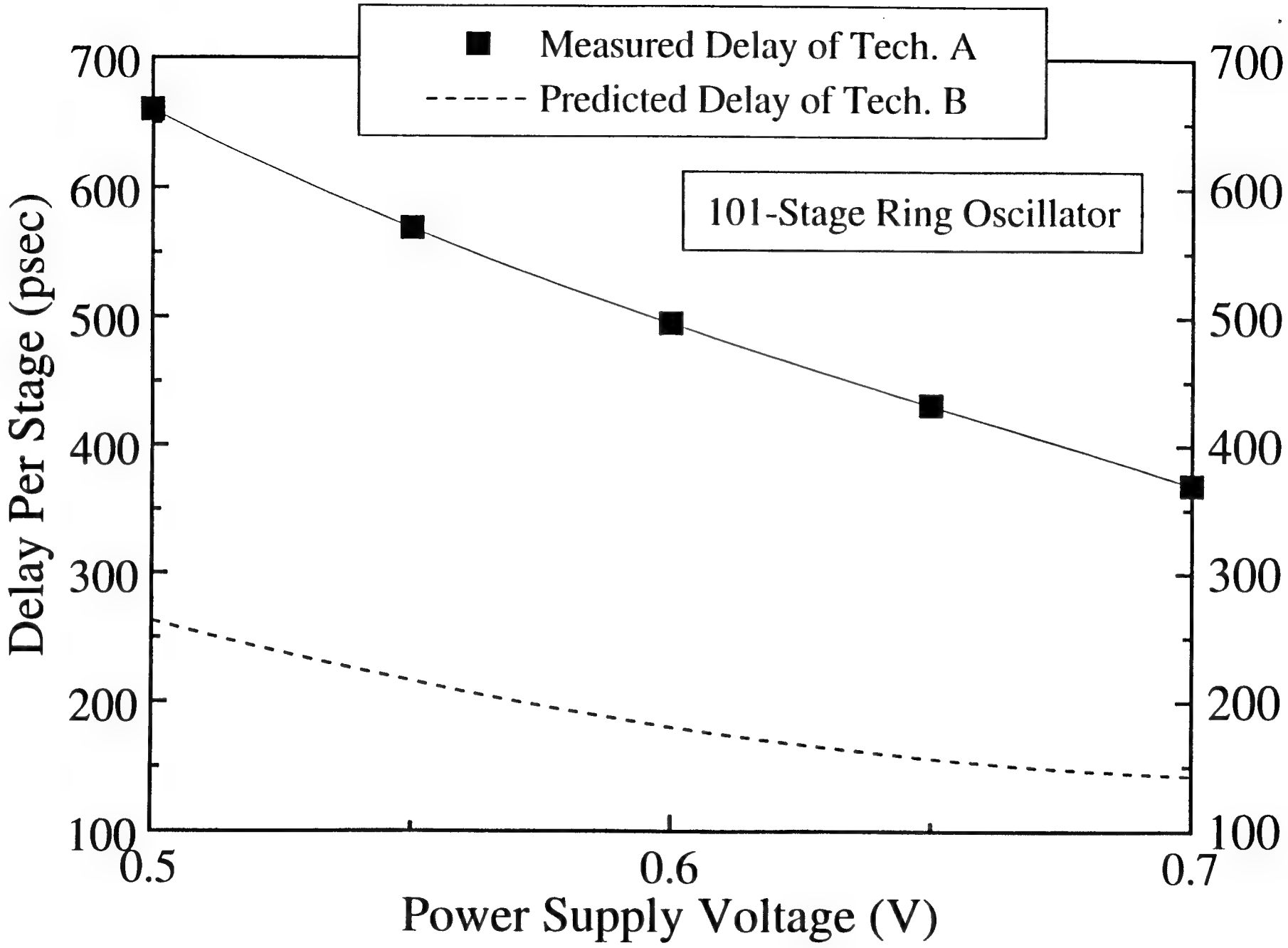

Threshold Voltage, V_t (V)
Body Bias, V_{bs} (V)
V_{to}
V_{tf}
V_{to}
V_{tf}
$V_{gs}=V_{bs}$
$V_{t,min}$
$V_{t,min}$
Technology-A
Technology-B
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

Fig.


Drain Current, VTMOS
Drain Current, Regular
Body Current
W=5μm, Technology-B
Drain Current (A)
Gate Voltage (V)
S=60mV/dec
S=81mV/dec
S=61mV/dec
S=80mV/dec
10-4
10-5
10-6
10-7
10-8
10-9
10-10
10-11
-0.7
-0.5
-0.3
-0.1
0.1
0.3
0.5
0.7


V_{gs} from 0.2V to 0.6V
in 0.1V steps
Body tied to the gate
Body grounded
Technology-B
L_{eff}=0.3μm
Drain Current (μA/μm)
80
70
60
50
40
30
20
10
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Drain Voltage (V)


Measured Delay of Tech. A
Predicted Delay of Tech. B
101-Stage Ring Oscillator
Delay Per Stage (psec)
700
600
500
400
300
200
100
0.5
0.6
0.7
Power Supply Voltage (V)

A Low-Barrier Body-Contact Scheme for SOI MOSFETs to Eliminate the Floating Body Effect

Mansun Chan, Zhi-Jian Ma, Fariborz Assaderaghi, Cuong T. Nguyen*,
Chenming Hu, and Ping K. Ko

Department of Electrical Engineering and Computer Science
University of California at Berkeley, CA 94720
* Stanford University, Stanford, California

INTRODUCTION

CMOS technology built on SOI substrate has many advantages over its bulk counterpart. The most often quoted are higher current drive, reduced parasitic capacitance, better device isolation and thus no latch up, and superior short-channel behaviors. It is also a simpler technology to develop and implement in the deep-submicron regime. However, the inherent floating body effect of SOI MOSFETs generates many problems, the most famous of which is the "kink" effect. A fully-depleted (FD) device reduces the kink, but, as this paper shows, does not solve many of the problems that limit its applicability, especially in analog circuits. We propose and demonstrate that a new simple Low-Barrier Body-Contact (LBBC) technology eliminates these problems.

FABRICATION PROCESS

Fig. 1 shows a NMOSFET with the new LBBC structure. Structure of the PMOSFET is similar. The fabrication process is a modified version of the conventional CMOS SOI process discribed in [1]. Only key steps of the NMOS process will be given here. After the gate definition step, a deep boron implant was performed with dose $1x10^{14}$ cm^{-2} and energy 60keV. This formed a moderately doped P region close to the buried oxide. Then shallow source/drain arsenic implant was performed with dose $3x10^{15}cm^{-2}$ and energy 25keV, followed by a shallow boron implant, which is the source/drain implant for PMOS in a CMOS process. A special implant mask was used to utilize this implantation step to give a P+ region right next to the N+ source of the NMOSFET. This P+ region is butted together with the N+ source with the same contact. The underlying P region formed by the tailored boron implant, either neutral or depleted, provides a low-barrier path for the holes generated by impact ionization to be collected through the butted P+ region. Ploeg proposed a conceptually similar dual source structure [2]. However the scheme depended on the Al spiking phenomenon, which is sensitive to process variation and incompatible with VLSI junction and contact techology (for example, silicidation).

DEVICE PERFORMANCE

The I-V characteristics of NMOSFETs and PMOSFETs fabricated with different technologies are shown in Fig. 2 and 3 respectively. The LBBC MOSFETs exhibit higher breakdown voltage especially at low V_g, and a very constant I_{dsat} which is free of kink. This shows that the LBBC is very effective in collecting the substrate current. Fig. 4

shows the collection efficiency of the LBBC compared with the bulk MOSFETs fabricated in the same lot. At low I_{sub}, the LBBC is capable of collecting the same amount of substrate current as that of the bulk. This substrate current collection scheme is in fact much more effective than the normal side-body contact scheme that sacrifices significant area.

The output resistance (R_{out}) of MOSFETs with LBBC is compared with conventional kink free FD MOSFETs (Fig. 5). The FD Structure essentially softened the kink, but did not actually eliminate it. So the resulting output resistance is very low, especially at low V_g due to threshold reduction as a result of DIBL and/or impact ionization (Fig. 7 and 8). As shown in figure 5, the R_{out} can be improved by at least one order of magnitude by using the LBBC technology. The negative resistance due to self-heating is suppressed by using a thick Si film (0.16μm) and thin buried oxide (0.11μm). Fig. 6 shows the voltage gain of LBBC MOSFETs and conventional FD MOSFETs. Much higher gain, which is important for analog applications, can be obtained with the LBBC technology.

The subthreshold characteristics of a 0.2μm LBBC and FD MOSFETs are shown in Fig. 7 and 8. The abnormally high subthreshold slope due to charging of floating body by impact ionization at high drain voltages [3] is completely removed. Thus, a much lower off-state leakage current and better gate control of drain current can be achieved for both PMOSFETs and NMOSFETs.

The flicker noise characteristics of the LBBC MOSFETs, Bulk MOSFETs and FD MOSFETs, another important consideration for analog application, are shown in Fig. 9 and 10. The LBBC MOSFETs shows a much lower flicker noise. In conventional FD MOSFETs, the floating body cannot sink any junction leakage and substrate current caused by hot-carrier effects, which can result in fluctuation of surface potential, which in turn modulates the channel carrier density. With the LBBC technology, the extra current can be sunk resulting in bulk like flicker noise level.

Fig. 11 and 12 compare the threshold voltage drop (ΔV_T) and subthreshold swing (S) shift due to short channel effect between the LBBC and FD MOSFETs. The slight random variation in V_T in the FD SOI is caused by variation of silicon film thickness which is not acceptable in low voltage digital circuits or high precision analog circuits. As can be seen, MOSFETs with the LBBC structure show an improved short channel behavior over conventional FD SOI MOSFETs.

CONCLUSION

The LBBC structure has been developed which can greatly improve SOI MOSFETs performance for digital and analog applications. The process only require 2 extra masks for a CMOS process and an insignificant amount of extra area. The LBBC is the most effective substrate currect collection scheme reported.

ACKOWLEDGEMENT

This project was supported by SRC under contract number 93-DC-324 and AFOSR/JSEP under contract number F49620-93-C0041.

REFERENCES:

[1] S. Parke et al., IEDM Tech. Dig., pp. 453-456, 1992

[2] E. Pleog et. al., IEDM Tech. Dig. pp. 337-340, 1992

[3] J. Fossum et. al., EDL-8, No.11, pp544, Nov. 1987

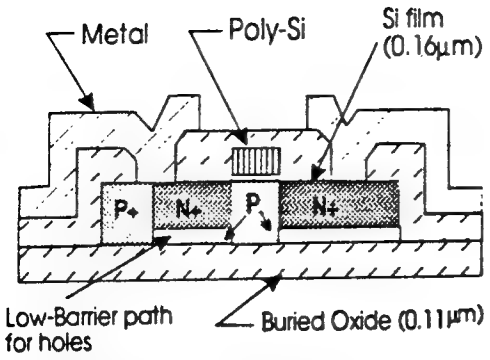


Fig. 1: NMOSFET with the Low-Barrier Body-Contact. The narrow P low-Barrier path was formed by a tailored $10^{14}cm^{-2}$ boron implant at source/drain implant step. PMOSFET fabricated has similar parameters

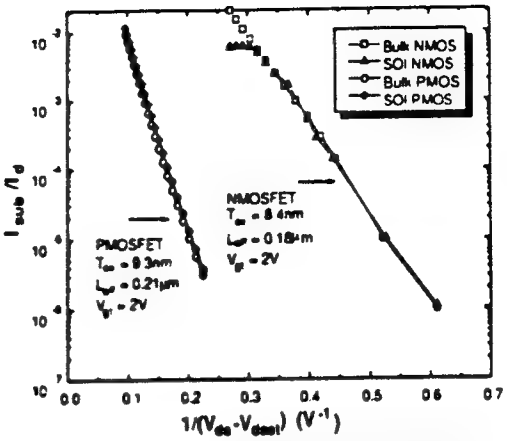


Fig. 4 Comparison of substrate current collection effectiveness between bulk MOSFETs and SOI MOSFETs with Low-Barrier Body-Contact

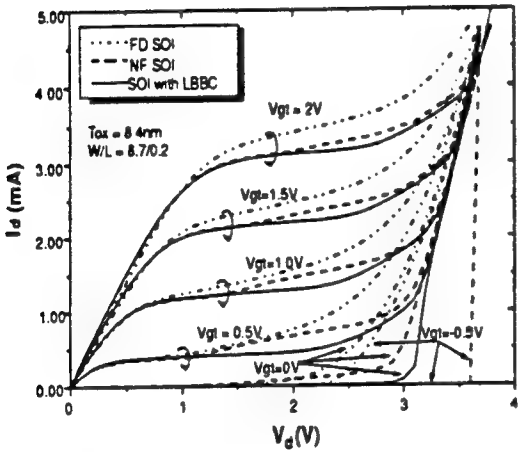


Fig. 2: I-V characteristics of NMOSFETs fabricated with different technologies

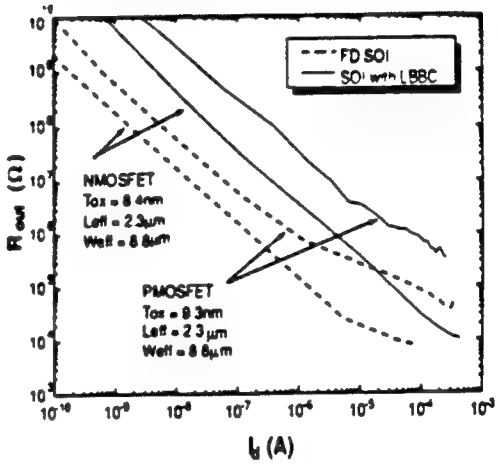


Fig. 5: Comparison of output resistance between SOI with Low-Barrier Body-Contact and fully-depleted SOI

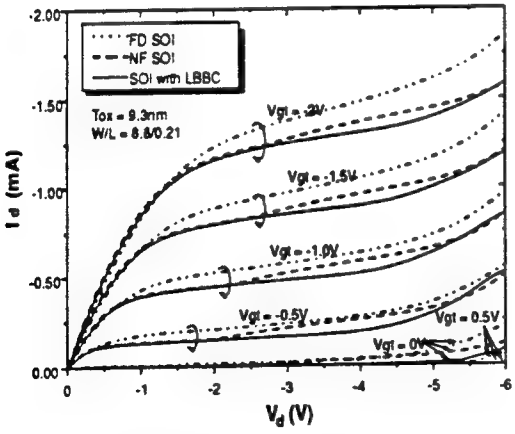


Fig. 3: I-V characteristics of PMOSFETs fabricated with different technologies

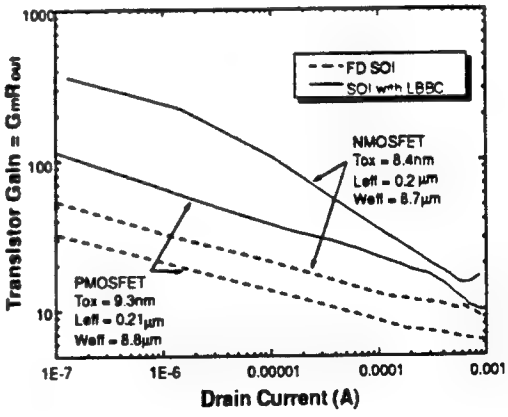


Fig. 6: Comparison of single transistor small signal voltage gain between fully-depleted SOI and SOI with Low-Barrier Body-Contact

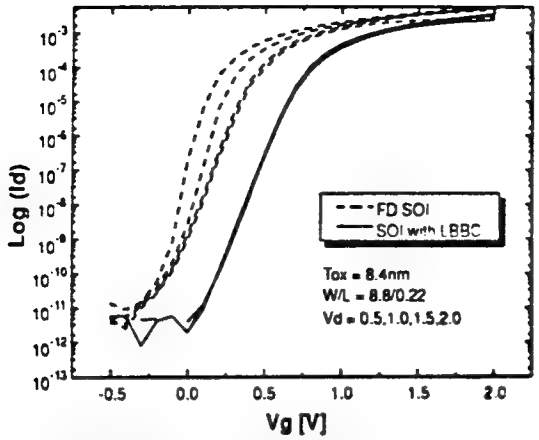


Fig. 7: Subthreshold characteristics of a fully-depleted SOI NMOSFET and SOI NMOSFET with Low-Barrier Body-Contact

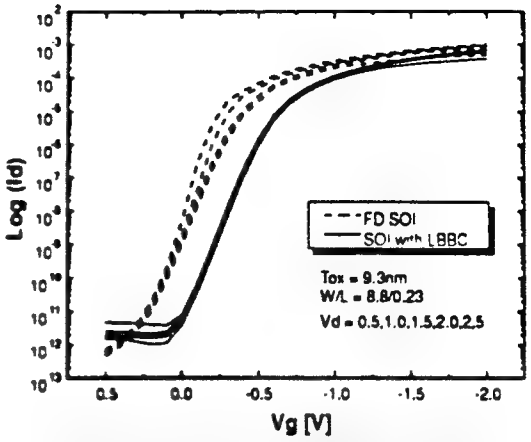


Fig. 8: Subthreshold characteristics of a fully-depleted SOI PMOSFET and SOI PMOSFET with Low-Barrier Body-Contact

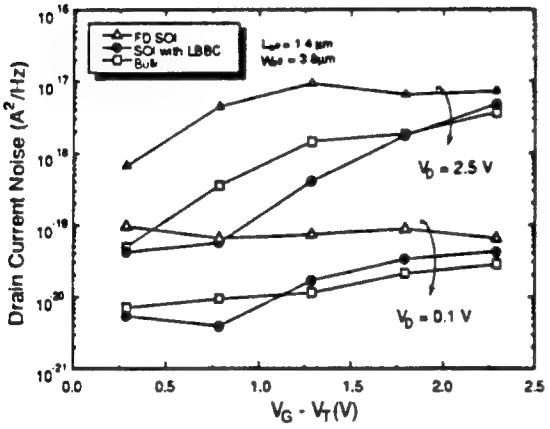


Fig. 9: Flicker noise characteristics measured on different NMOSFET structures

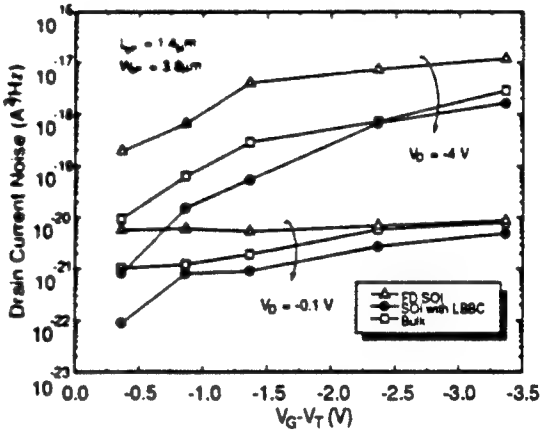


Fig. 10: Flicker noise characteristics measured on different PMOSFET structures

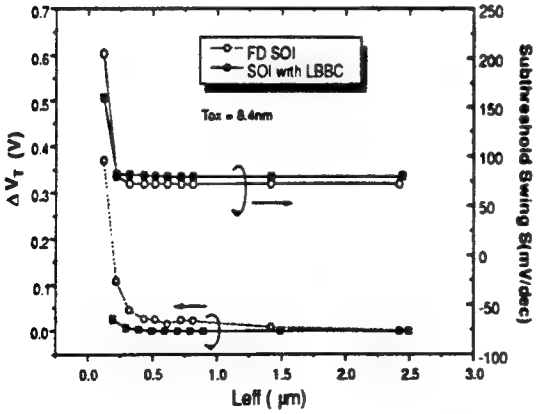


Fig. 11: NMOSFET threshold voltage shift (ΔV) and subthreshold swing (S) versus effect channel length at V_{ds}= 0.05V

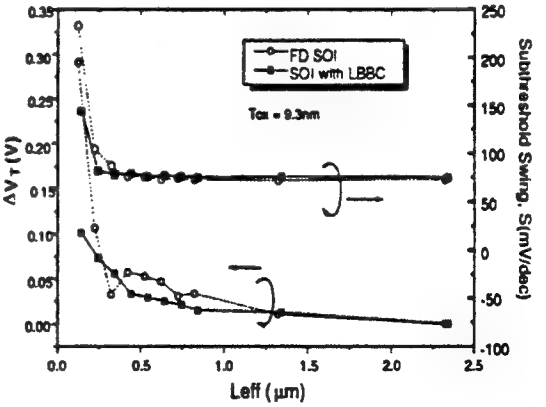


Fig. 12: PMOSFET threshold voltage shift (ΔV) and subthreshold swing (S) versus effect channel length at V_{ds}= -0.05V

RECESS CHANNEL STRUCTURE FOR REDUCING SOURCE/DRAIN SERIES RESISTANCE IN ULTRA-THIN SOI MOSFET (Revised, MS#741)

Mansun Chan, Fariborz Assaderaghi, Stephen A. Parke*, Chenming Hu and Ping K. Ko
Department of Electrical Engineering and Computer Science,
University of California at Berkeley, CA 94720

*Stephen A. Parke is with the IBM Corporation, East Fishkill, NY

Abstract

A new Recess-Channel SOI (RCSOI) Technology has been developed for fabricating ultra-thin SOI MOSFETs with low source/drain series resistance. Thin-film fully-depleted SOI MOSFETs with channel film thickness of 72nm have been fabricated with the RCSOI technology. The new structure demonstrated a 70% reduction in source/drain series resistance compared with conventional processes. In the deep-submicron regime, more than 80% improvement in saturation drain current and transconductance over conventional devices was achieved using the RCSOI technology. The new technology would also facilitate the use of silicide for further reducing the series resistance.

INTRODUCTION

Fully-depleted (FD) MOSFETs fabricated on ultra-thin silicon-on-insulator (SOI) films have received significant attention for integrated circuit applications due to reduced parasitic capacitance, simple device isolation, better short channel behavior and radiation hardness [1]. However, as silicon film thickness (t_{si}) is reduced, source/drain series resistance (R_{sd}) of SOI MOSFETs increases, which in turn significantly reduces the current drive and speed response in the deep-submicron regime [2]. Silicide technology, which is used to reduce R_{sd} in bulk MOSFETs, is difficult to apply to ultra-thin film SOI wafers. For example, Schottky diode behavior has been observed for nMOSFETs in some SAlicide SOI processes [3], most likely caused by additional lateral migration of the silicide into the source/drain junction due to the rapid lateral consumption of the finite volume of silicon. Also, high R_{sd} may still result after silicidation, as reported in [4]. In this letter, we propose and demonstrate that a simple Recess-Channel SOI (RCSOI) technology can significantly reduce R_{sd} of thin-film FD SOI MOSFETs. The process is much simpler and more robust than other technologies for the same purpose using selective epi-silicon [5] and selective CVD Tungsten [6], especially when t_{si} at the channel is below 100nm.

FABRICATION PROCESS

The fabrication process is a modified version of the conventional CMOS SOI MESA process described in [7]. The key steps are shown in Fig. 1. The starting wafers have silicon film thickness of 195nm. A LOCOS process was performed at the channel region forming 180nm of oxide, which consumes 85nm of silicon. The oxide was then wet etched giving a silicon film of 110nm at the channel. Because it was not a self-aligned process, a margin of 0.3μm, limited by layer-to-layer registration, was given to both sides of the gate. Mesa active islands (MESA) were created by plasma-etching a nitride/oxide/silicon stack stopping at buried oxide. Next, a 100nm oxide was grown on the MESA sidewalls to prevent low-V_T edge devices and gate oxide defects at the MESA corners. Threshold implant was then performed, resulting in concentration of 1-

$3x10^{17}cm^{-3}$ in the silicon film. Gate oxide of 8.6nm was grown, followed by the deposition of 270nm of polysilicon. Doping of the poly-silicon gate was realized by a 25-keV, $5x10^{15}cm^{-2}$ boron, and 50keV, $5x10^{15}cm^{-2}$ phosphorus implant giving P+ gate for pMOSFETs and N+ gate for nMOSFETs. Photoresist-ashing process [8] was then performed to give deep-submicron transistors with effective channel length as small as 0.1μm. Source/drain implant with As for nMOSFETs and BF_2 for pMOSFETs was done using an implant energy of 30-keV and a dose of $3x10^{15}cm^{-2}$. The final cross-section is also shown in Fig. 1.

DEVICE PERFORMANCE

The RCSOI MOSFETs demonstrated here have final silicon film thickness of 165nm at the source/drain region and 72nm at the channel after subsequent oxidation and etching. The contact-to-gate spacing is about 1.5μm. Table 1 and Fig. 2 summarizes the performance of MOSFETs with effective channel length (L_{eff}) of 0.3μm, fabricated using different technologies. The high R_{sd} results from conventional process reduces the saturation drain current (I_{dsat}) and saturation transconductance (g_{msat}) of deep-submicron MOSFETs significantly. The saturation voltage (V_{dsat}) in the presence of high R_{sd} is also much higher, thus preventing the use of deep-submicron MOSFETs for low voltage applications. The RCSOI technology is capable of reducing the R_{sd} by a factor of 3 at this channel film thickness. This factor is expected to be larger when the silicon film is thinner. 80% improvement in I_{dsat} and g_{msat} over conventional devices has been achieved using the RCSOI technology. The V_{dsat} of the new devices also reduced by 45% for nMOS and 35% for pMOS. Lower V_{dsat} is the cause of lower breakdown voltage in RCSOI MOSFET (Fig. 2). Note that the thick-film SOI MOSFETs used for comparison here are non-fully-depleted (NFD), which have lower intrinsic I_{dsat} and g_{msat} due to the substrate charge effect [9]. But its lower R_{sd} makes its performance comparable with the FD SOI MOSFETs.

The measured I_{dsat} versus L_{eff} of different MOSFETs are shown in Fig. 3. Decreasing channel length aggravates the reduction of I_{dsat} (and similarly, g_{msat}) in conventional SOI MOSFETs

because the debiasing effect of R_{sd} becomes stronger as current increases, thus diminishing the advantages of scaling. With the RCSOI technology, 85% of the intrinsic I_{dsat} (assuming $R_{sd} = 0$) can be achieved versus less than 50% achieved by conventional process at L_{eff} below 0.5μm. The impact of reduction in t_{si} on R_{sd} and I_{dsat} is shown in Fig. 4. Note that the t_{si} is the measured [10] film thickness at the channel, not the thickness at the source/drain region. Due to the re-oxidation after the source/drain implant and over-etch in contact opening, the silicon film on the source/drain region is expected to be thinner. The very high series resistance in 42nm silicon film maybe caused by contact problem when making contact to the ultra-thin silicon film. Such contact problem can be eliminated by the RCSOI technology, thus arbitrary thin SOI MOSFETs can be fabricated. Also silicide technology may be used in conjunction with the Recess-Channel technology to further reduce the source/drain series resistance. The thicker silicon film in the source/drain region provides more silicon for the formation of silicide, making silicide process much easier to apply.

The only potential drawback of the RCSOI technology is the non-self-aligned nature of the process which may result in asymmetric devices characteristics. However, this kind of behavior was not observed in our measurement.

CONCLUSIONS

A new Recess-Channel SOI technology has been developed. It significantly reduces R_{sd}, thus increasing the current drive and the transistor gain in deep-submicron SOI MOSFETs. This technology is potentially very useful for fabricating high performance ultra-thin SOI MOSFETs with arbitrary silicon film thickness. The process is compatible with most of the existing CMOS processes, including silicidation for further reducing the source/drain series resistance.

ACKNOWLEDGMENT

This project was supported by SRC and AFOSR/JSEP under contract number F49620-93-C0041.

References

[1] H. Miki, T. Ohmameuda, M. Kumon, K. Asada, T. Sugano, Y. Omura, K. Izumi, and T. Sakai, "Subfemtojoule deep submicrometer-gate CMOS built in ultra-thin Si film on SIMOX substrates," *IEEE trans. Electron Devices*, vol. ED-38, pp. 373-376, Feb. 1991.

[2] M. Jeng, J. E. Chung, P. K. Ko, and C. Hu, "The effects of source/drain resistance on deep submicrometer device performance," *IEEE Trans. Electron Devices*, vol. ED-37, pp. 2408-2410, Nov. 1990.

[3] S. Tyson and R. Gallegos, "Salicided source/drain considerations on UTF SIMOX," 1991 IEEE International SOI Conference, pp. 66-67.

[4] N. Kistler, E. V. Ploeg, J. Woo, and J. Plummer, "Sub-quarter-micrometer CMOS on ultrathin SOI," *IEEE Electron Device Lett.*, vol. EDL-13, 1992, pp. 235-237.

[5] J. R. Rfiester, M. Woo, J. T. Fitch and J. Schmidt, "Reverse Elevated Source/Drain (RESD) MOSFET for Deep Submicron CMOS," *IEDM Tech. Dig.*, 1992, pp. 885-888.

[6] D. Hisamoto, K. Nakamura, M. Saito, N. Kobayashi, S. Kimura, R. Nagai, T. Nishida, and E. Takeda, "Ultra-Thin SOI CMOS with Selective CVD Tungsten for Low Resistance Source and Drain," *IEDM Tech. Dig.*, 1992, pp. 829-832.

[7] S. Parke, F. Assaderaghi, J. Chen, J. King, C. Hu, and P. Ko, "A Versatile, SOI BiCMOS Technology with Complementary Lateral BJT's," *IEDM Tech. Dig.*, 1992, pp. 453-456.

[8] J. Chung, M. Jeng, J. E. Moon, A. T. Wu, T. Y. Chan, P. K. Ko, and C. Hu, "Deep-submicrometer MOS devices fabrication using a photoresist-ashing technique," *IEEE Electron Device Lett.*, vol. EDL-9, pp. 186-188, 1988.

[9] J. F. Fossum and S. Krishnan, "Current-Drive Enhancement Limited by Carrier Velocity Saturation in Deep-Submicrometer Fully Depleted SOI MOSFET's," *IEEE Trans. Electron Devices*, vol. ED-40, pp. 457-459, Feb. 1993.

[10] J. Chen, R. Solomon, T. Y. Chan, P. K. Ko, and C. Hu, "A CV Technique for Measuring Thin SOI Film Thickness", *IEEE Electron Device Lett.*, vol-12, No. 8, pp. 453-455, Aug. 1991.

Figure Captions

Fig. 1: Key steps of the RC SOI process and the final cross-section of the device

Table 1: A summary of device performance. The MOSFETs demonstrated have T_{ox} = 8.6nm, L_{eff} = 0.3μm. I_{dsat} and V_{dsat} are measured at $V_{gt} = V_{gs} - V_{th}$ = 1.5V and G_{msat} is measured at V_{ds} = 2V. Subthreshold swing (S) is measured at V_{ds}=0.1V. The film thicknesses are 72nm for thin-film, 165nm for thick-film. The RCSOI structure has 72nm at the channel and 165nm at the source drain region.

Fig. 2: I-V characteristics of a conventional thin-film SOI nMOSFET and a nMOSFET with Recess-Channel structure.

Fig. 3: Measured I_{dsat} versus L_{eff} of conventional SOI MOSFETs and MOSFETs with RCSOI structure.

Fig. 4: Experimentally measured R_{sd} and I_{dsat} versus silicon film thickness. The MOSFETs used have L_{eff} = 0.3μm, T_{ox} = 8.6nm. I_{dsat} is measured at V_{gt}=1.5V absolute value. The I_{dsat} of RCSOI MOSFETs is also shown as a comparison.

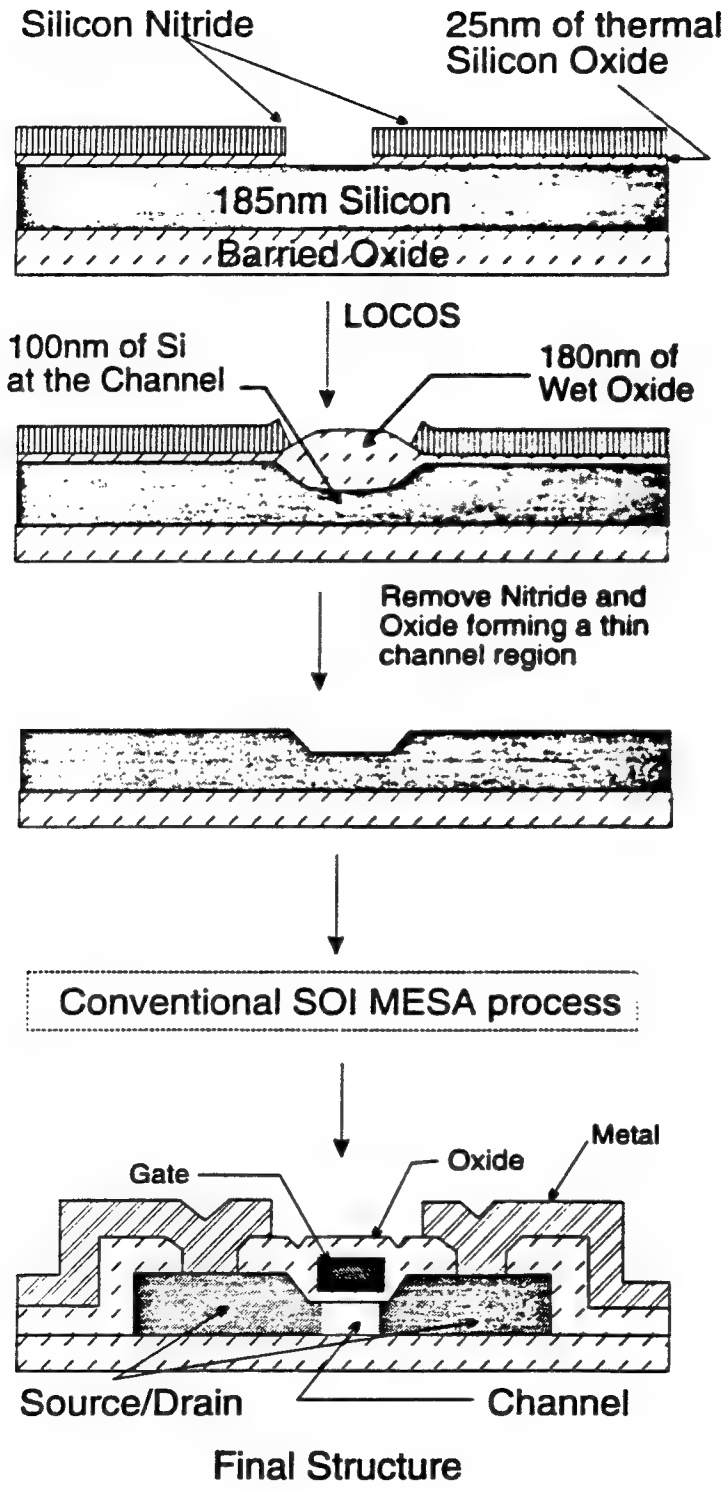

Silicon Nitride
25nm of thermal Silicon Oxide
185nm Silicon
Barried Oxide
LOCOS
100nm of Si at the Channel
180nm of Wet Oxide
Remove Nitride and Oxide forming a thin channel region
Conventional SOI MESA process
Gate
Oxide
Metal
Source/Drain
Channel
Final Structure

Fig. 1

	R_{sd}(Ω–μm)	I_{dsat} (mA/μm)	V_{dsat} (V)	G_{msat} (S/mm)	S (mV/dec)
NMOS					
Thin-film	3420	0.17	1.31	0.14	68.2
Thick-film	930	0.33	0.70	0.23	75.2
RC	1044	0.32	0.71	0.24	69.4
PMOS					
Thin-film	6880	0.07	1.28	0.058	67.3
Thick-film	1710	0.13	0.82	0.097	74.1
RC	2145	0.13	0.82	0.099	67.9

Table 1

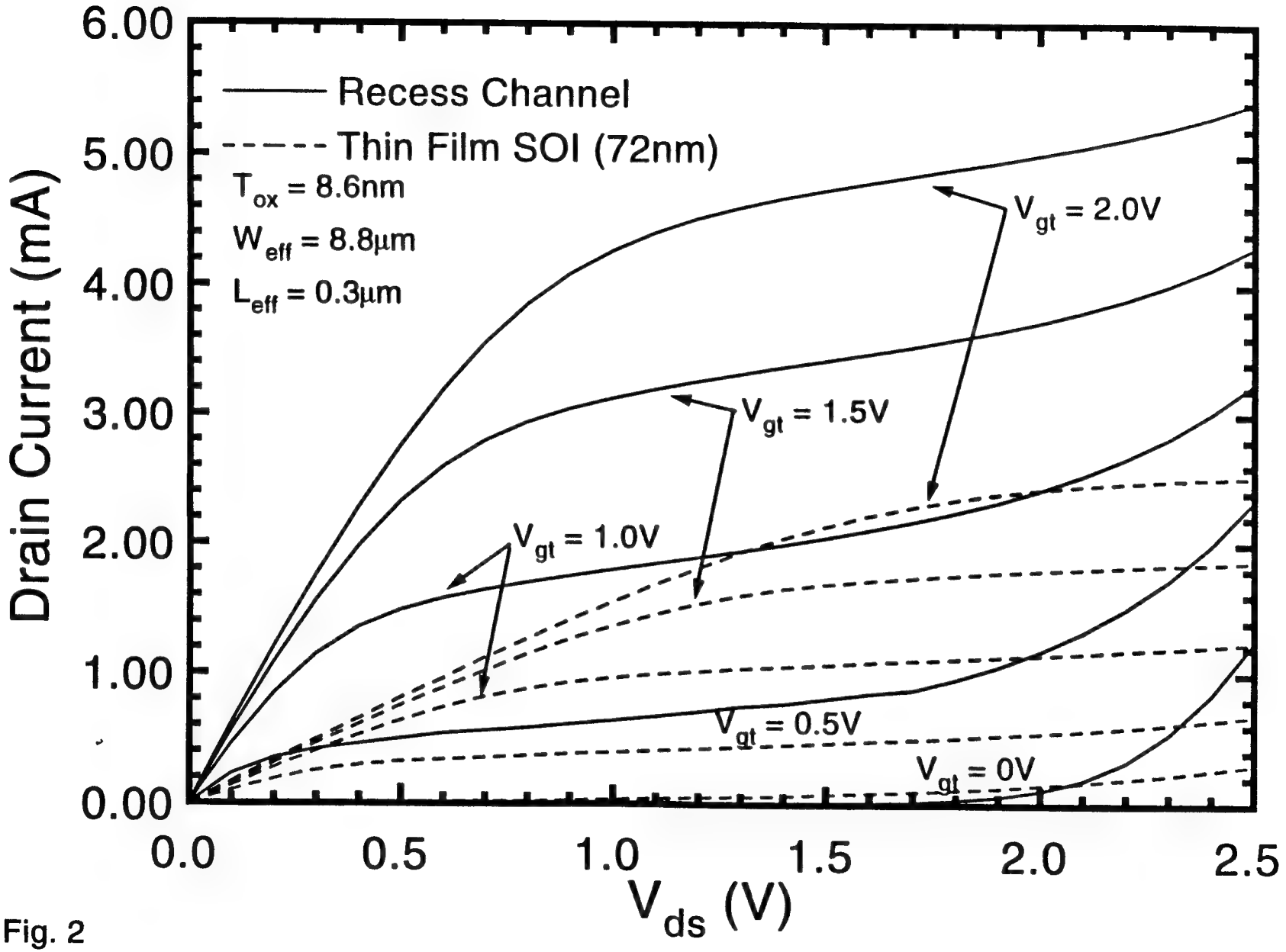

6.00
5.00
4.00
3.00
2.00
1.00
0.00
Drain Current (mA)
Recess Channel
Thin Film SOI (72nm)
T_{ox} = 8.6nm
W_{eff} = 8.8μm
L_{eff} = 0.3μm
V_{gt} = 2.0V
V_{gt} = 1.5V
V_{gt} = 1.0V
V_{gt} = 0.5V
V_{gt} = 0V
0.0
0.5
1.0
1.5
2.0
2.5
V_{ds} (V)

Fig. 2

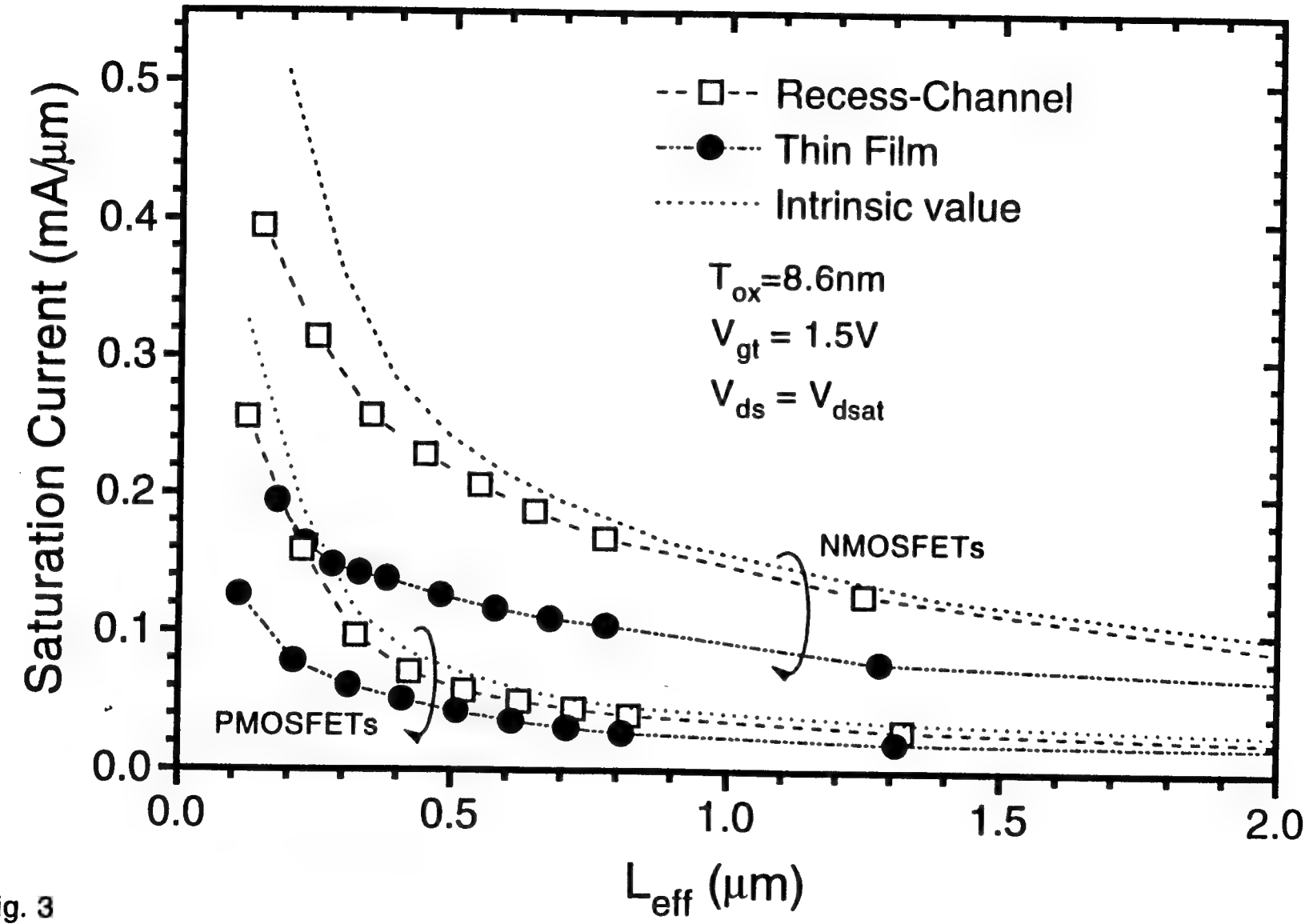

Saturation Current (mA/μm)
0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.5
1.0
1.5
2.0
L_{eff} (μm)
Recess-Channel
Thin Film
Intrinsic value
T_{ox}=8.6nm
V_{gt} = 1.5V
V_{ds} = V_{dsat}
NMOSFETs
PMOSFETs

Fig. 3

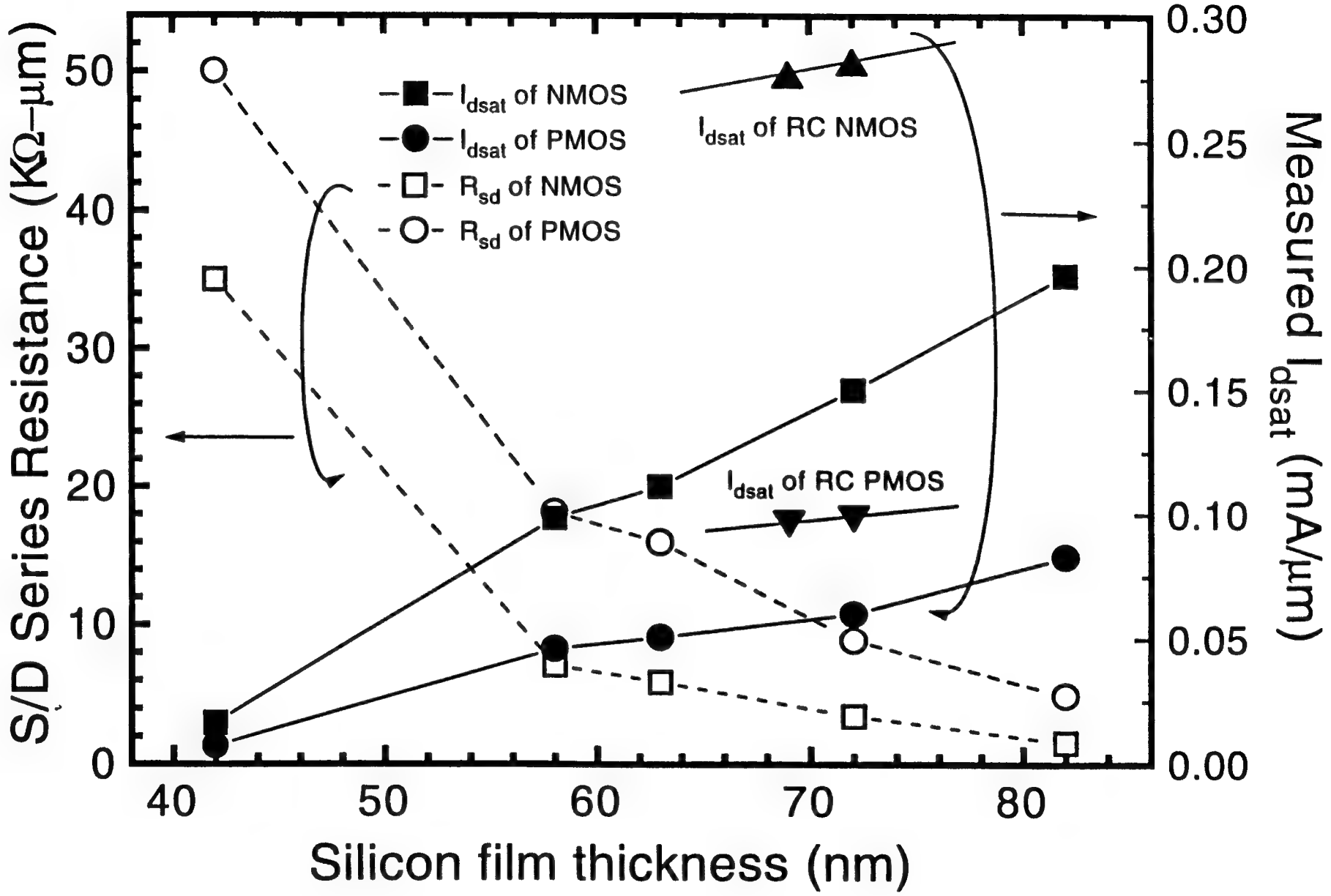

S/D Series Resistance (KΩ-μm)
Measured I_{dsat} (mA/μm)
Silicon film thickness (nm)
I_{dsat} of NMOS
I_{dsat} of PMOS
R_{sd} of NMOS
R_{sd} of PMOS
I_{dsat} of RC NMOS
I_{dsat} of RC PMOS
0
10
20
30
40
50
0.00
0.05
0.10
0.15
0.20
0.25
0.30
40
50
60
70
80

Fig. 4

Comparison of ESD Protection Capability of SOI and BULK CMOS Output Buffers

Mansun Chan, Selina S. Yuen, Zhi-Jian Ma, Kelvin Y. Hui,
Ping K. Ko and Chenming Hu

Department of Electrical Engineering and Computer Science
University of California at Berkeley,
Berkeley, CA 94720

ABSTRACT

ESD protection capability of SOI CMOS output buffers has been studied with Human Body Model (HBM) stresses of both positive and negative polarity. Experimental results show that the ESD discharge current is absorbed by the NMOSFET alone. Unlike bulk technologies where the bi-directional ESD failure voltages are limited by positive polarity stresses, SOI circuits display more serious reliability problem in handling negative ESD discharge current. Bulk NMOS output buffers fabricated on the substrate of the same SOI wafers, after etching away the buried oxide, have been used to compare the ESD protection capability between bulk and SOI technologies. The ESD voltage sustained by these "bulk" NMOS buffers is about twice the voltage sustained by conventional SOI NMOS buffers. This scheme is proposed as an alternative ESD protection for SOI circuits. The effectiveness of ESD resistant design strategies developed in bulk-substrate technologies when transferred to SOI circuits is also discussed in this paper.

INTRODUCTION

As VLSI circuits are becoming more performance driven, Silicon-On-Insulator (SOI) CMOS technologies have become very attractive. By dielectrically isolating circuit elements, SOI technologies eliminate transistor latch-up and provide reduced junction capacitance. Such reduction in parasitic capacitance allows IC's to operate at much higher circuit speeds than conventional bulk-substrate silicon IC's with the same device dimensions. Because of the better short-channel behavior, higher circuit density, and simpler fabrication process, SOI technologies show great potential to become the low-cost mainstream production technologies [1].

With the rapid advancement of SOI technology, electrostatic discharge (ESD) susceptibility becomes one of the major reliability issues. However, very little attention has been paid to ESD phenomena for SOI circuits. In bulk-substrate technologies, good ESD protection levels have been demonstrated by using NMOS/CMOS output buffers [2,3]. However, most of the protection schemes developed for bulk may not be compatible with SOI structures. For example, the use of thick-field-oxide devices becomes impractical on SOI wafers. Large-area low-series-resistance (vertical) PN junctions are not available either, as the silicon film is usually thinner than 150nm. Several ESD protection schemes designed for SOI circuits have been proposed, which use additional circuits constructed with diodes and polysilicon resistors [4]. These solutions consume large silicon area, introduce large delays, and are far from adequate.

In this paper, the ESD susceptibility of submicron SOI NMOS/CMOS output buffers is studied with HBM stresses. The failure mechanisms of these buffers are investigated to provide an understanding of SOI ESD phenomena. The impact of different design parameters such as gate-to-contact spacing, silicon film thickness (T_{Si}) and effective channel width (W_{eff}) on ESD susceptibility during HBM stresses are also presented. The results are compared with NMOS buffers which have similar physical structures and fabrication processes to reveal the effectiveness of improving SOI ESD performance by design strategies. And finally, an alternative ESD protection scheme of fabricating the output buffers on the substrate of SOI wafers is discussed.

EXPERIMENTATION

To study the ESD phenomena of SOI circuits, non-optimized deep-submicron MOS buffers with 250μm effective channel width have been fabricated. The layout utilized the 'finger structure' as shown in Fig. 1 to achieve a more uniform current density [5]. The silicon film thickness and the buried oxide thickness (T_{box}) are 163nm and 100nm respectively. All these transistors are fabricated on SIMOX (Separation by IMplanted OXygen) wafers with MESA isolation process [6]. A thin gate oxide of 8nm is used to suppress short channel effects in the deep-submicron regime, where SOI technologies show

CH3332-4/94/0000-0292$01.00 © 1994 IEEE IRPS

superior performance over bulk technologies. The effective channel lengths (L_{eff}) ranged from 0.3μm to 1.5μm and the gate-to-contact spacing is 2μm. Because the breakdown characteristics of a NMOSFET can be very different with body floating and body grounded, MOSFETs with special body contacts [7] are also tested in the study.

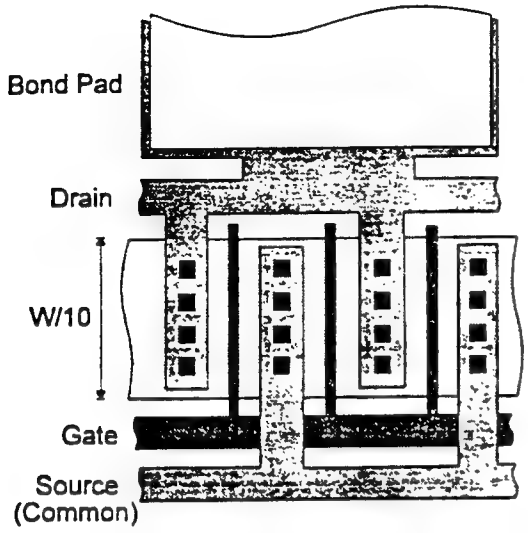


Fig. 1. Ladder structure used in NMOS output buffer Layout

The bulk NMOS buffers used for comparison are fabricated on the substrate of the same SOI wafer after etching away the buried oxide. The structure is shown in Fig. 2. This allows us to compare the ESD protection capability of SOI and bulk-substrate output buffers with similar fabrication processes and physical structures.

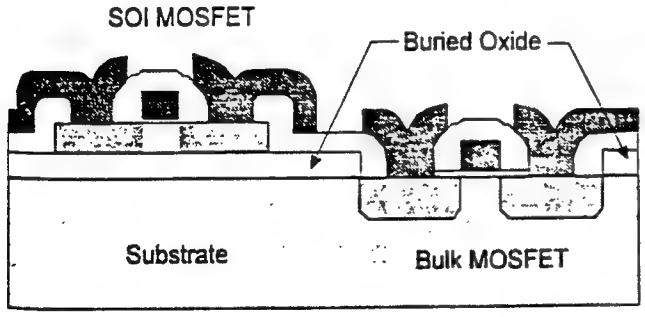


Fig. 2. Structure of a SOI NMOSFET and a BULK NMOSFET being tested

The transistors are stressed according to the Human Body Model (HBM) as specified in Mil-Std 883C Method 3015.7. Eight to Ten transistors are stressed with 3 pulses per stress level to obtain a statistical distribution of ESD failure voltages at each stress condition. A maximum leakage current of 100nA for drain voltage ranging from -0.5V to 3V is chosen as the failure criterion.

POSITIVE POLARITY HBM DISCHARGE OF NMOS OUTPUT BUFFERS

During positive polarity HBM discharge, NMOSFET operating in the bipolar breakdown/snapback mode is usually used as a clamping device [9,10]. Fig. 3 and 4 shows the breakdown/snapback characteristics of the bulk and SOI NMOSFETs respectively.

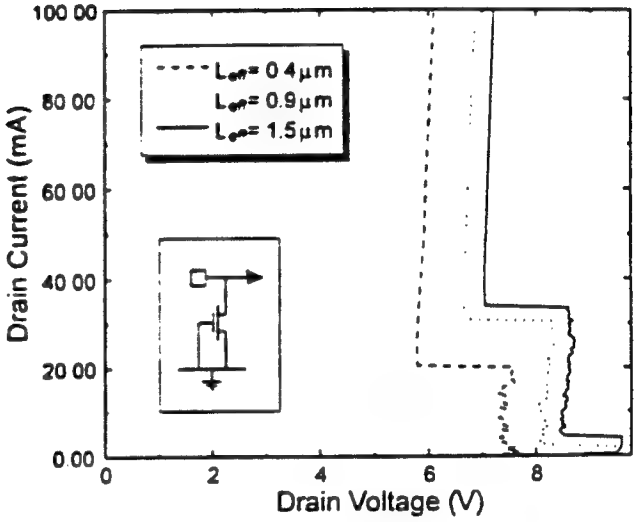


Fig. 3. Breakdown/snapback characteristics of bulk NMOSFETs with different effective channel lengths

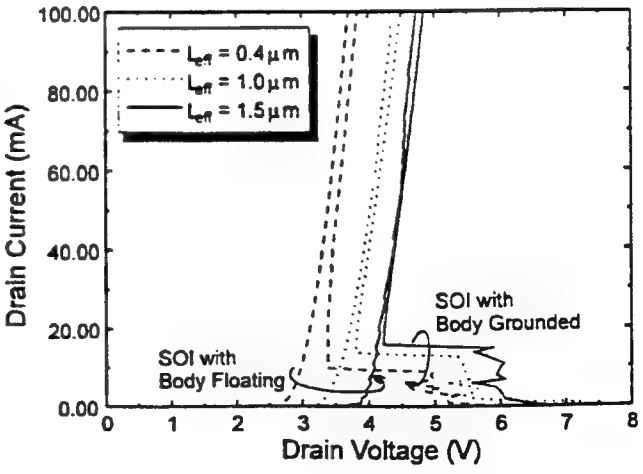


Fig. 4. Breakdown/snapback characteristics of SOI NMOSFETs with different effective channel lengths. Measurements are done for both body floating and body grounded.

A double snapback breakdown behavior is observed in the bulk NMOSFETs and body-grounded SOI NMOSFETs, and this double snapback phenomenon has been claimed to give reasonable ESD protection [10,11,12]. No snapback behavior can be observed when the body of the SOI NMOSFET is left floating because the breakdown voltage BV_{CEO} is lower than the snapback voltage. The second snapback of SOI NMOSFETs occurs at a smaller drain current with lower holding voltage when compared with the bulk case. Thus, SOI NMOS buffers are more efficient in clamping ESD discharge voltages. However, as the second snapback is believed to be caused by current localization due to the negative resistance coefficient of silicon beyond a critical temperature [13,14], the lower second snapback current also indicates more serious Joule heating taking place in the SOI NMOSFETs. It is reasonable since the heat sink capability of silicon is much higher than the heat sink capability of SiO_2, which completely surrounds the active silicon island in the SOI NMOS buffers. If we take the ratio of the snapback voltage-current products of bulk (Fig. 3) and SOI (Fig. 4) MOSFET, as the inverse ratio of the device thermal resistance, the SOI device thermal resistance is about 2-4 time that of bulk device. The resistance in the second breakdown regime (inverse of slope) is also larger for SOI devices (Fig. 4) due to higher current density than bulk devices (Fig. 3). This is deleterious for ESD relibility.

In actual positive polarity HBM ESD discharge, the average ESD voltages sustained by SOI NMOSFETs are about 580V for a 250μm wide device, only half of that sustained by bulk NMOSFETs which average to 1020V. The box plot in Fig. 5 shows the median, interquartile ranges, and the extremums of HBM ESD voltage levels withstood. No significant difference is observed between the body-floating and the body-grounded case, in agreement with the simlarity of holding voltages in Fig. 4.

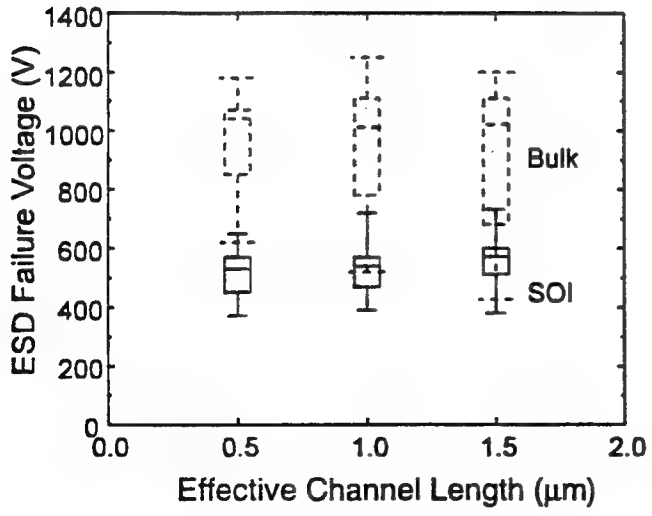


Fig. 5. ESD failure voltage of bulk and SOI NMOSFETs under positive ESD HBM stress

The failure mode for the SOI NMOSFETs was found to be a short among the gate, drain and substrate which can be seen as in Fig. 6. A similar failure mode is observed in the bulk NMOSFET. The failures are believed to be caused by Joule heating resulting in a thermal runaway condition during second breakdown [15]. The temperature at the drain junction is high enough to cause silicon melting and ejection through the thin gate oxide [16], thus causing a short between the gate, drain and substrate.

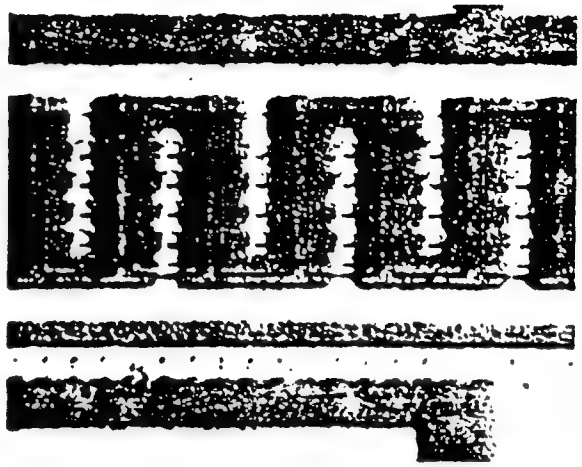


Fig. 6. Example of SOI NMOSFET ESD failure under positive ESD HBM stress

It is interesting to observe that reducing the effective channel length of the bulk NMOSFET increases the mean and reduces the spread at the lower end of ESD failure voltages, but it produces the opposite trend for SOI NMOSFETs. As the amount of ESD stresses sustained is believed to be determined by the silicon temperature, the effect of increasing the gate length for SOI NMOSFETs has, according to the observed ESD test results, a stronger impact on increasing the heat sink capability, which more than compensates for the adverse effect of increasing snapback holding voltage and series resistance.

NEGATIVE POLARITY HBM DISCHARGE OF NMOS OUTPUT BUFFERS

The failure voltages of SOI and bulk NMOSFETs under negative HBM ESD stress are shown in Fig. 7. In bulk NMOSFETs, the negative polarity HBM discharge pulses are absorbed by the large drain to substrate forward biased diode. This allows the transistor to sustain higher (about 300V in our case) negative discharge voltage compared with the case of positive polarity discharge. However, in SOI technologies, large-area vertical PN junctions are not available and the discharge current path is restricted to the thin active silicon film

which is usually thinner than 150nm. Due to the reduced level of parasitic bipolar action during negative discharge, the negative HBM discharge current is clamped by the NMOSFET operating in the so-called transistor-diode mode. Since the series resistance of the NMOSFET in this operating mode is relatively high, together with the high current density restricted in the thin-film, serious local heating results in a much lower negative HBM discharge current or voltage.

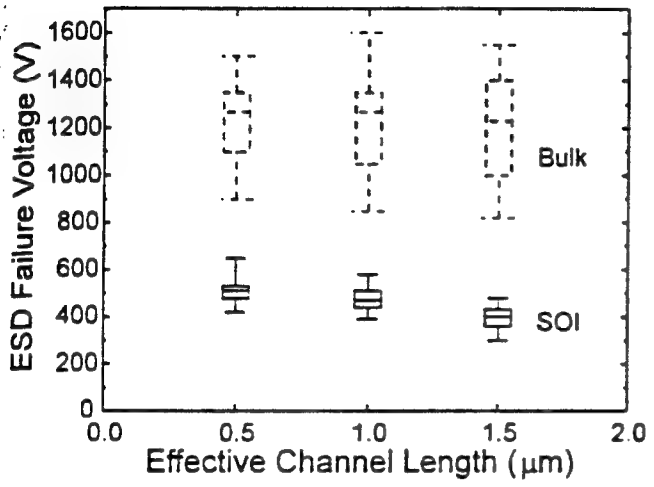


Fig. 7. ESD failure voltage of bulk and SOI NMOSFETs under negative ESD HBM stress

Experimental results shows that the average ESD discharge voltage sustained by SOI NMOSFETs during negative discharge is about 10% lower compared with the positive HBM discharge case at L_{eff}=0.5μm. The main failure mode is again a short among the gate, drain and substrate for both the SOI and bulk NMOSFETs. But in some SOI NMOSFETs, gate oxide rupture is observed, which is caused by the potential difference between drain and gate due to the high series resistance. Again no observable difference can be found between the body-ground and body-floating case. As the effective channel length increases, the negative ESD voltage sustained by the SOI NMOSFET decreases and more gate oxide rupture is observed. It is not surprising since the series resistance of the transistor increases with channel length, resulting in a more serious heating and higher potential across the gate oxide.

From the above results, we see that the bi-directional ESD HBM discharge susceptibility is limited by negative polarity discharge. Since higher negative ESD failure voltages can be attained by reducing the channel length, the gate length should be kept small to improve the overall ESD reliability of SOI NMOS buffers.

ESD Performance of SOI CMOS Output Buffers

In conventional bulk CMOS output buffers, positive polarity ESD HBM discharge limits the ESD reliability. And it has been reported that the parasitic devices in a bulk CMOS well process can play a significant role during an ESD event by providing additional current paths for positive stress current with respect to either V_{DD} or V_{SS} [5,17,18]. These do not apply to SOI CMOS technologies because no well is available in the thin film. Besides, the ESD HBM failure voltage of SOI NMOSFETs is limited by negative polarity discharge as indicated in the previous sections. Since breakdown/snapback does not occur in PMOSFET until very high voltages (15V in our case), the addition of the PMOSFET cannot improve the performance of the NMOSFET during negative ESD HBM discharge. On the other hand, because the snapback voltage of NMOSFET is low, the PMOSFET cannot help in the positive ESD HBM discharge either.

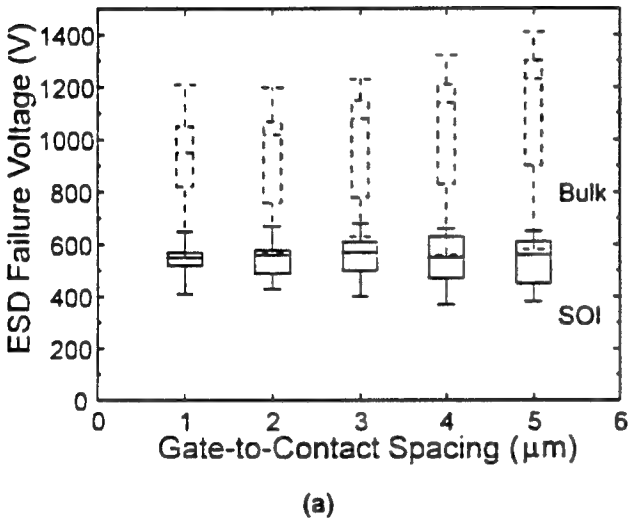
Fig. 8 shows the average ESD failure voltages of SOI CMOS buffers with and without p-channel device under both positive and negative HBM stresses. No significant difference can be observed whether the PMOSFET is present or not. This confirms that ESD protection in SOI CMOS buffers is provided by the NMOSFET alone.

	Failure Voltage under positive HBM Stress	Failure Voltage under negative HBM Stress
With both P-ch and N-ch Devices	232V	187V
With N-ch Device Only	224V	192V

Fig. 8. Comparison of ESD performance of a CMOS output buffer device with and without the p-channel device. The transistors under test has dimension W/L = 20μm/1μm for the NMOSFET and W/L = 30μm/1μm for the PMOSFET. They all have T_{ox} = 7nm, T_{Si} = 150nm, T_{box} = 400nm. 6 transistors are stressed to obtain the mean of the distribution.

Other Device Parameters Related to ESD Protection Capability

Many researchers have shown that increasing the gate-to-contact spacing in abrupt junctions NMOS/CMOS output buffers can improve the ESD protection capability [2,5,17,19]. By varying the gate-to-contact spacing from 1μm to 5μm, the average ESD failure voltage of bulk NMOSFETs under positive HBM stress increase by about 200V as shown in Fig. 9 (a). But it has no observable effect in the positive HBM ESD failure voltages of SOI NMOSFETs.



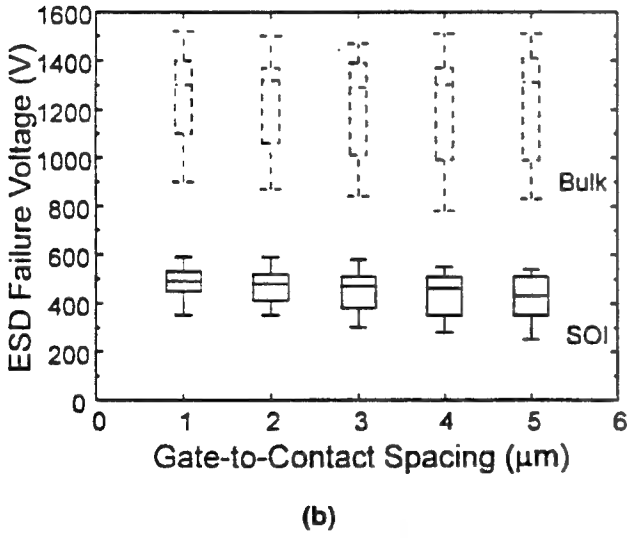


*Fig. 9. ESD failure voltage of NMOSFETs with different Gate-to-Contact spacing under **(a)** positive ESD HBM stress, and **(b)** negative ESD HBM stress. The transistors have L_{eff} = 1μm, W_{eff} = 250μm and T_{ox} = 8nm*

It can be explained by the three dimensional nature of bulk MOSFETs, in which case the substrate is capable of sinking the heat generated along the current path from the contact to the gate. However, due to the presence of an insulating buried oxide, the heat generated in a SOI MOSFET can only flow laterally, resulting in roughly the same temperature at the drain/channel junction regardless of gate-to-contact spacings. Thus a similar ESD failure voltage is obtained.

Furthermore, increasing the gate-to-contact spacing results in a higher series resistance during negative discharge, which causes more power dissipation in the transistor. Thus, a larger gate-to-contact spacing even lowers the negative ESD failure voltages of SOI MOSFETs as shown in Fig. 9 (b). Therefore, due to bi-directional ESD stress consideration, gate-to-contact spacing should be kept small.

Silicon film thickness is another important design parameter in SOI circuits in determining the operating modes of the transistors [20]. At the same power dissipated in the transistor, the silicon temperature increases with decreasing silicon film thickness because of the reduction of heat capacity in smaller silicon volume [21]. More serious local heating may also result because of the higher series resistance and higher current density confined in the thin film. As a result, the ESD performance will be worse as the silicon film thickness is scaled down. Fig. 10 shows the average ESD failure voltage as a function of silicon film thickness which confirm the prediction.

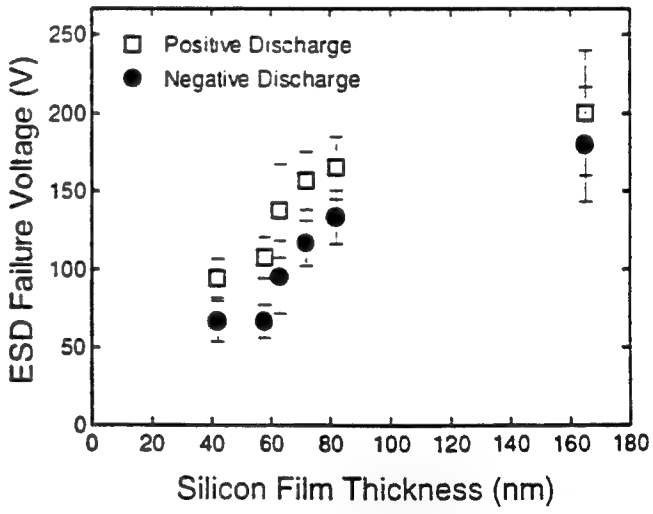
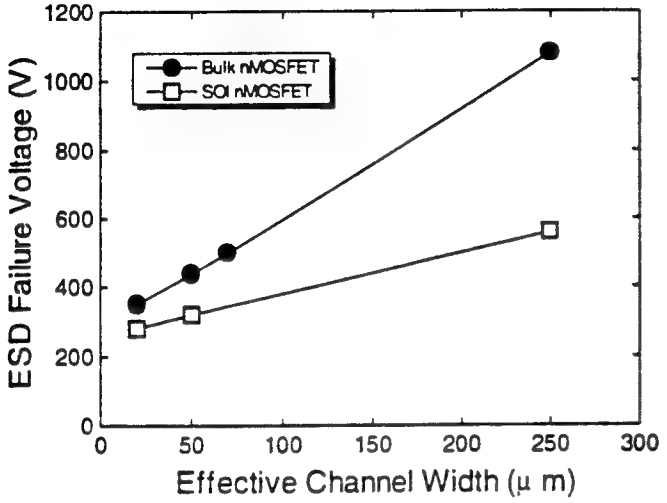
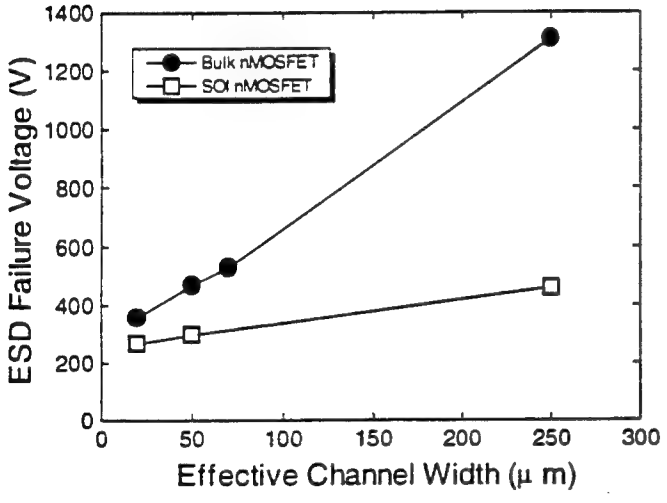


Fig. 10. Mean and standard deviation of ESD failure voltages of SOI NMOSFETs verses different silicon film thickness under both positive and negative ESD stress. The transistors have L_{eff} = 1μm, W_{eff} = 20μm and T_{ox} = 7nm. Silicon film thickness is measured by the method described in [22]

By increasing the channel width of the MOSFET, ESD performance improves accordingly. However, in SOI circuits, the ESD failure voltage does not increase as much with increasing device width compared with the bulk technologies, which is illustrated in Fig. 11. This prevents the attainment of good ESD protection by simply enlarging the device width as is routinely done for bulk IC ESD protection.



(a)



(b)

Fig. 11. *ESD failure voltage versus device width for NMOSFETs for SOI and bulk technologies. The transistors have L_{eff} = 1μm and T_{ox} = 8nm. The results are for Human Body Model Test under* ***(a)*** *positive stress, and* ***(b)*** *negative stress.*

In general, most of the strategies used for improving ESD performance in bulk technologies do not apply to SOI circuits. ESD protection schemes have to be re-optimized to provide better SOI ESD reliability.

CONCLUSION

In this paper, the ESD protection capability of non-optimized submicron ultra-thin gate oxide SOI and bulk CMOS buffers are studied and compared using Human Body Model stresses. Results show that the ESD voltages sustained by SOI NMOS/CMOS buffers are only about 55% of those achieved by the bulk technology. This is mainly attributed to the poor heat dissipation due to the insulating buried-oxide layer, causing higher temperature in the silicon film during an ESD event. Due to the absence of large vertical PN diodes, the ESD bi-directional stress is limited by negative polarity stress pulses. To obtain the maximum bi-directional HBM ESD protection level, the channel length, should be kept minimal. Our study also shows that most of the methods developed in bulk technologies to improve ESD performance do not work as well in SOI circuits, thus different strategies should be investigated.

As an alternative ESD protection scheme, we propose to design ESD protection circuits on the silicon substrate through openings in the buried oxide, created by an extra masking step. With the CMOS protection circuits, one may choose to build both NMOSFET and PMOSFET or only the NMOSFET in the substrate (while keeping the PMOSFET in the Si film) in order to simplify the process. As the above results show, this protection scheme is capable of improving the ESD performance by 100% and allows most of the ESD protection schemes developed for bulk technologies to be directly transferred to the SOI technologies.

ACKNOWLEDGMENT

This project was supported by SRC under contract 93-DC-008, AFOSR/JSEP under contract F49620-93-C0014 and a grant from the ESD Association. The authors would also like to acknowledge Joseph King and Y. C. Shih of UC Berkeley for the valuable discussions on this work and help in preparing the figures and slides.

REFERENCES

[1] Laura Peters, "SOI Takes Over Where Silicon Leaves Off", Semiconductor International, March 1993, pp. 48-51

[2] R. Rountree and C. Hutchins, "NMOS Protection Circuitry", IEEE Trans. Electron Devices, ED-32, No. 5, May 1985

[3] C. Duvvury and R. N. Rountree, "A Synthesis of ESD Input Protection Scheme", Proceedings of the 1991 EOS/ESD Symposium, pp. 88-97

[4] W. Palumbo and M. Dugan, "Design and characterization of input protection networks for CMOS/SOS applications", Proceedings of the 1986 EOS/ESD Symposium, pp. 182-187

[5] C. Duvvury, R. N. Rountree, Y. Fong, and R. A. McPhee, "ESD Phenomena and Protection Issues in CMOS Output Buffers", Proceedings of 1987 IEEE/IRPS, pp. 174

[6] S. Parke, F. Assaderaghi, J, Chen, J. King, C. Hu, and P. K. Ko, "A Versatile, SOI BiCMOS Technology with Complementary Lateral BJT's", in IEDM Tech. Dig., 1992 pp. 453-456.

[7] Z. J. Ma, H. J. Wann, M. Chan, J. King, Y. C. Cheng, P. K. Ko, and C. Hu, "Characterization of Hot-Carrier Effects in Thin-Film Fully-Depleted SOI MOSFETs", Proceedings of 1994 IEEE/IRPS, 1994.

[8] A. R. Pelella, and H. Domingos, "A Design Methodology for ESD Protection networks", in Proceedings of the 7th EOS/ESD Symposium, pp. 20-40, 1985

[9] C. Duvvury, and R. Rountree, "Output ESD Protection Techniques for Advanced CMOS Processes", in Proceedings of the 12th EOS/ESD Symposium, pp. 206-211, 1988.

[10] K. Verhaege, G. Groeseneken, J. P. Colinge, and H. E. Maes, "Analysis of Snapback in SOI NMOSFETs and its Use for an SOI ESD Protection Circuit", in Proceedings of the IEEE SOI Conference, pp. 140-141, 1992.

[11] K. Verhaege, G. Groeseneken, J.-P. Colinge, and H. E. Maes, "Double Snapback in SOI NMOSFETs and its Application for SOI ESD Protection", IEEE Electron Device Lett., Vol. 14, No. 7, July 1993, pp. 326-328.

[12] J. S. T. Huang, "A Model for Double Snapback Phenomena in N Channel SOI MOSFETs", in Proc. 1993 IEEE SOI Conf., pp. 122-123.

[13] W. B. Smith, D. H. Pontius, and P. P. Budenstein, "Second Breakdown and Damage in Junction Devices", IEEE Trans. on Electron Devices, vol. ED-20, 1973, pp. 731-744

[14] D. R. Alexander, "Electrical Overstress Failure Modeling for Bipolar Semiconductor Components", IEEE Trans. Comp. Hybrids Manuf. Technol., vol. CHMT-1, 1978, pp. 345-353

[15] A. Amerasekera, L. Roozendall, J. Bruines, and F. Kuper, "Characterization and Modeling of Second Breakdown in NMOSTs for the Extraction of ESD-Related Process and Design Parameters", IEEE Trans. on Electron Devices, vol. ED-38, No. 9, Sept. 1991, pp. 2161-2168.

[16] M. A. Bridgwood, "Breakdown Mechanisms in MOS Capacitors Following Electrical Overstress", Proceedings of the 1986 EOS/ESD Symposium, pp. 200-207

[17] N. Khurana, T. J. Maloney and W. Yeh, "ESD on CHMOS Devices: Equivalent Circuits, Physical Models and Failure Mechanisms", Proceedings of 1985 IEEE/IRPS, pp. 212

[18] K. L. Chen, G. Giles, and D. B. Scott, "Electrostatic Discharge Protection for One Micron CMOS Devices and Circuits", in IEDM Tech. Digest, 1986.

[19] T. Polgreen and A. Chatterjee, "Improving the ESD Failure Threshold of Silicided NMOS Output Buffers by Ensuring Uniform Flow", IEEE Trans. Electron Device, ED-39, No. 2, Feb. 1992

[20] J. P. Colinge, "Silicon-on-Insulator Technology: Materials to VLSI", edited by Kluwer Academic Publishers, Boston 1991.

[21] L. T. Su, K. E. Goodson, D. A. Antoniadis, M. I. Flik, and J. E. Chung, "Measurement and Modeling of Self-Heating Effects in SOI NMOSFETs", in IEDM Tech. Dig., 1992 pp. 357-360.

[22] J. Chen, R. Solomon, T. Y. Chan, P. K. Ko, and C. Hu, "A CV Technique for Measuring Thin SOI Film Thickness", IEEE Electron Device Lett., vol-12, No. 8, pp. 453-455, Aug. 1991.

Intermediate View Reconstruction for Three-Dimensional Scenes

Nelson L. Chang and Avideh Zakhor

Department of Electrical Engineering and Computer Sciences
University of California, Berkeley, CA 94720 USA
e-mail: nlachang@robotics.Berkeley.EDU, avz@eecs.Berkeley.EDU

Abstract

This paper is concerned with three-dimensional scene reconstruction algorithms. The goal is to implement a system in which the video sequence obtained from movement of a camera around a 3-D object can be used to reconstruct arbitrary intermediate views not directly recorded by the camera. This problem has been solved for the case in which the object is defined "synthetically" in the computer. The main difference between our approach and the existing ones is that we use real objects and real video cameras, rather than mathematically defined objects typically used in computer graphics applications. Our approach is to "scan" a 3-D object by translating a camera across it, construct the resulting depth map, and reconstruct arbitrary views based on computing the transformation between the camera locations. An interesting application of this system occurs when the shape of a 3-D object needs to be transmitted to a remote location. The idea here is to build a 3-D representation of the object from its recorded video signature at the transmitter. At the receiver, the viewer's head is tracked in such a way as to update the view on a stereoscopic display, thus creating a 3-D impression of the object at the receiver end. We will show simulation results on 3-D intermediate view reconstruction.

1 Introduction

With the development of faster memory and graphics hardware, there has been increased interest in the development of virtual environments. Researchers have been investigating the representation of mathematically defined or "synthetic" objects. The result has led to "virtual" worlds which create a 3-D impression of the object to the viewer as he/she moves around.

Our interest is to extend this idea to real 3-D objects. Many applications of such an environment immediately come to mind: the designer who wants to show some prospective clients on the other coast her design; the real-estate agent who displays houses by having interested parties "walk-through" a simulation; and the surgeon who studies a 3-D simulation before attempting the real surgery. With special hardware such as a stereoscopic display and a head tracking device, the viewer is able to gain the sense of 3-D where the system updates the display according to the movement of the viewer's head. Thus, the goal is to devise a compact representation which sufficiently captures the 3-D information of the real object.

There are several possible approaches to this scene representation/reconstruction problem. Certainly, the easiest approach would be to capture a very large number of views of a given object and store these images off-line in a huge image database for later reconstruction. While this solution would provide high quality reconstructed images, it is neither practical nor efficient, for it requires large amounts of memory and it does not exploit the inherent 3-D geometry of the scene. Another possibility is to model the object by mathematical formulae and store this reduced set of information. This approach saves in storage, but would require a complex analysis of the object, and the accuracy of the reconstructions can be obtained only with models having a large number of degrees of freedom. Instead, we desire an approach that would provide high quality reconstructions and yet would not need a great deal of memory.

In this paper, we consider an approach where a camera has scanned a given stationary object along several pre-specified trajectories. From each of these sequences, we recover depth information at certain

locations, and use this information, along with the corresponding intensities, to generate fairly accurate reconstructions.

In Section 2, we describe the algorithms for deriving the compact representation and scene reconstruction. Section 3 contains the experimental results on a particular object. Finally, we conclude in Section 4 with a discussion of our approach.

2 Description of the Scene Reconstruction Algorithms

In order to construct a system which enables users to visualize objects, two issues should be addressed: the representation of objects and their reconstruction from an arbitrary view.

2.1 Derivation of the Compact Representation

To derive a compact representation of a 3-D object, we must first devise a method for acquiring the necessary information. We propose to capture several video sequences by scanning a camera along a number of trajectories with known geometries. An example of a rectangular scanning pattern with four linear trajectories is shown in Figure 1. In this figure, the camera is assumed to translate across each trajectory A, B, C, and D. We may also scan along another set of trajectories at a different elevation. By doing so, more 3-D information of the object is captured. Note that the second set of trajectories are not necessarily mere translations along the y direction of the first set of trajectories; translation along the z-component may also be present. Our goal is to extract sufficient yet compact information from these scanned frames so that we may reconstruct an arbitrary view of the object anywhere along the trajectories.

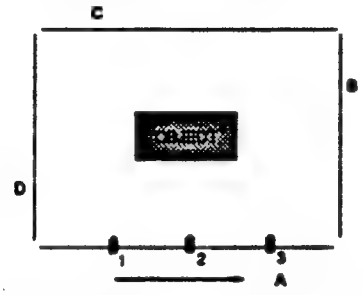


Figure 1: *An example of "scan" geometry along four linear trajectories A, B, C, and D.*

One possible representation of the 3-D scene consists of the depth and intensity at selected frames of each trajectory, e.g. at locations 1, 2, and 3 of trajectory A in Figure 1. These selected frames are referred to as reference frames. We believe that from this set of data, we can reconstruct intermediate views of the object at arbitrary points on the trajectories. Assuming that the reference frames from each trajectory have already been selected, the steps for deriving the representation are as follows:

1. *Derive an initial estimate of depth for each frame relative to the nearest reference frame.* We first solve the correspondence problem by matching features between each frame and the nearest reference frame. There are several approaches to address this problem; for simplicity, we choose to perform a simple block matching search for each feature. The depth is then simply inversely related to the disparity between matched pixels, i.e. if Δx is the disparity between a feature in the reference frame and the current frame, then the depth for those points is approximately equal to $1/\Delta x$. Thus, for each frame, we generate a frame of depths for every pixel, so-called depth map.

2. *Compute and equalize scaling factor between depth maps.* In the previous step, we find the depths to within a scale factor. It is quite possible that the scale factors among the depth maps all differ. Before we may go on to the next step, we must first determine the factor of each depth map and then scale them with respect to one depth map.

To determine the scale factor, we find the region of pixels with a particular minimum depth in one depth map and then determine the depths of the corresponding points in a second depth map. We may use the feature correspondences from the first step to aid in identifying the corresponding points. As described in [1], we solve a linear regression problem where the depths of the pixels in the first depth map $\vec{z}_1$ are matched to those in a second depth map $\vec{z}_2$. If α is the scale factor, then the estimate for α is given by

$$\alpha = \frac{\sum_{i \in K} \vec{z}_1}{\sum_{i \in K} \vec{z}_2}$$

where the set K consists of the pixels for which depth is defined in both $\vec{z}_1$ and $\vec{z}_2$. Once the factor is determined for every depth map, then we scale each depth map by its factor so that they are all equalized with respect to the same depth map.

3. *Combine depth maps to obtain accurate depth information at each reference frame.* The depth map associated with each reference frame is determined by the neighboring equalised depth maps. For every point, we examine the depth at the corresponding point in each of the neighbors, remove the outliers which are greater than one standard deviation from the mean, and combine the rest by weighted sum to generate a single value. We then have a collection of accurate depth maps for each reference frame.

4. *Estimate camera motion between reference frames.* We identify edges in the reference frames and use them to determine camera motion. For details of the approach, see [1, 2, 3].

The compact representation of the object then consists of the set of depth maps, intensities, and motion parameters for each designated reference frame.

2.2 Reconstruction of an Intermediate View

Once we have generated the compact representation for a particular 3-D object, we may choose to reconstruct an intermediate view of the object at a specified point on one of the trajectories. Assuming that the relative position and orientation in space of the desired intermediate frame are known, the steps for reconstruction are as follows:

1. *Choose the appropriate reference frame(s) to use.* From the relative position and orientation of the desired frame, we should decide which reference frame(s) to use. One way of deciding is to use the reference frames which have the smallest motion parameters relative to the intermediate frame.

   Another consideration is the number of reference frames. If the intermediate frame is very close to one of the reference frames in the database, then we may choose to use only that reference frame for reconstruction, referred to as unilateral reconstruction. It is also possible that the particular view falls along a linear path between two reference frames. In this case, using both reference frames in bilateral reconstruction may be better. Finally, the view may lie within a region defined by four reference frames as in the case of two pairs of reference frames at two different elevations; quadrilateral reconstruction may be the best choice in this case.

2. *Generate estimates of the intermediate view by applying motion parameters to each reference frame.* We are assuming implicitly that the relative motion from the intermediate view to each of the chosen reference frames is known. The notion of applying motion parameters to a frame has been addressed in conventional computer vision literature [4, 5]. If (X_1, Y_1) are the initial image coordinates, (X_2, Y_2) are the final coordinates after motion, and z is the depth at the given point, then

$$X_2 = \frac{(r_1 X_1 + r_2 Y_1 + r_3)z + \Delta x}{(r_7 X_1 + r_8 Y_1 + r_9)z + \Delta z}$$

$$Y_2 = \frac{(r_4 X_1 + r_5 Y_1 + r_6)z + \Delta y}{(r_7 X_1 + r_8 Y_1 + r_9)z + \Delta z}$$

   where the parameters $r_1, r_2, \ldots, r_9$ are rotation parameters and Δx, Δy, Δz are translation parameters as described in [3].

3. *Combine the estimates of the previous step from each reference frame and interpolate the data to the nearest grid points.* Once we compute the estimates of the desired view with respect to each of the chosen reference frames, we must decide how to combine this data to generate the appropriate reconstruction. Furthermore, it is quite possible after the previous step that the estimates do not coincide with the sampling grid, so we must also ensure that the data is interpolated to the grid points.

   We propose the following technique for interpolation. Suppose there are N reference frames for reconstruction and suppose that there exists at least one data point near the given pixel (i, j). Then, the intensity value of the pixel on the sampling grid is given by

$$I(i,j) = \sum_{n=1}^{N} w_n \left(\frac{\sum_{k_n \in A_{ij}} d_{k_n} I_{k_n}}{\sum_{k_n \in A_{ij}} d_{k_n}} \right)$$

   where A_{ij} is a $2\Delta \times 2\Delta$ region centered around pixel (i, j), Δ is the spacing on the sampling grid, d_{k_n} is the distance of the $k_n th$ point in frame n with intensity I_{k_n} from the (i, j) pixel, and w_n is the weight for the data from reference frame n. Generally, these weights depend on the location with respect to the reference frames.

   It is possible that there exists no points in a given $2\Delta \times 2\Delta$ region, creating what we refer to as "holes." This condition occurs for de-occluded regions, that is, areas which are uncovered after the occluding object moves. In this case, we grow the area A_{ij} out to a $2m\Delta \times 2m\Delta$ region, where m is the smallest value for which a point falls within the area A_{ij}, i.e. the region is no longer a hole. Once we find such an area, we then use the above interpolation formula to find the intensity value at the grid point (i, j).

   Another refinement which improves reconstruction is to consider depth when interpolating. When an object moves in a frame, we would like the pixels of the object to have more weight than those from the background occupying the same region. One possible solution is to place more value on those pixels with smaller depth (closer) and less on those with larger depth (farther away). The result is that pixels with more motion will dominate over those that tend to be stationary.

3 Experimental Results

We shall now examine some results using the techniques described above. To generate these results, we scan an orange-juice container along four linear trajectories. The result consists of a 40-frame sequence per trajectory. A typical frame, e.g. frame #8 from view 2, is shown in Figure 2.a. For each of the four sequences, frames 8, 20, and 32 have been chosen to be the reference frames.

Following the algorithm outlined in Section 2.1, we derive the depth maps for each of the twelve reference frames. An example of the depth map corresponding to frame #8 is shown in Figure 2.b. The depth map has been heavily quantized to enable visualization. As can be seen, the area corresponding to the container is generally lighter than the background indicating that it is closer to the camera than any other object in the scene. The depth maps associated with the other reference frames are similar. It is interesting to note that certain patches of the background appear to be light in color. This is because of the mismatches associated with solving correspondence for areas of constant intensity.

Using the depths as well as the intensities of the reference frames, we reconstruct the views along the four trajectories according to the algorithm in Section 2.2 for bilinear reconstruction. For analysis, the *worst* reconstructions from two of the four trajectories are shown in Figure 3. Note that a measure of error is not included since the original sequence might not contain the exact frames corresponding to the motion parameters used to generate the intermediate view.

Figure 3.a is the intermediate view 40% and 60% of the translational motion between references frame 8 and 20, respectively, for trajectory 2. We observe that the overall quality of the image is good, due to the very accurate depth map for frame 8 shown in Figure 2.b. Some errors occur along the left edge of the container. The reason is that the depth map corresponding to frame 8 gives points immediately to the left of the box, which should be background, more depth than the background, causing the resulting interpolated image to be erroneous. This can also be attributed to not de-emphasizing enough the intensities of points from the stationary background compared with those in the moving container. In our algorithm, the depth

is used to determine which pixels to de-emphasize; it appears that the depth information may be too noisy for this purpose.

The reconstruction in Figure 3.b shows the view 40% from frame 8 for trajectory 3. Most of the artifacts occur in the left part of the stool and the upper-left portion of the orange juice container. We believe that a lot of spurious matches near these constant-intensity regions cause the depth maps to be inaccurate, and thus lowering the quality of the reconstructions.

4 Discussion

We have proposed an approach for representing and reconstructing stationary 3-D objects. The reconstructions in the last section seem to indicate that this approach is very promising. Many of the artifacts occur at the boundaries of the objects in the scene, and they stem primarily from inaccurate depths at these points. The block-matching technique we employ is simple yet reasonably adequate to solve the correspondence problem, however it is not the most optimum. Problems occur in the background when there is little movement. Other techniques such as hierarchical searches and gradient methods may be able to improve the results. Another interesting approach is to estimate motion and depth simultaneously [6].

Future work in this area includes examining the optimum number of reference frames to fully capture an object. A more complete analysis must be performed in order to determine what the scope of a single reference frame is, or conversely, what the optimum set of reference frames to compactly represent a given object is. In addition, a real-time implementation of the reconstruction algorithm would expedite the development of a virtual environment. Using a stereoscopic display and head tracking device, we will be able to simulate such a system by reconstructing an arbitrary view of an object in real time as the user moves his/her head. The area of scene reconstruction and its application to virtual environments seems very fertile and this research serves a good starting point.

Acknowledgments

This work was supported by NSF-PYI grant MIP–9057466, ONR Young investigator award N00014-92-J-1732, Joint Services Electronics Program (JSEP) contract F49620-90-0029, and Sun Microsystems.

References

[1] A. Zakhor and F. Lari, "Edge-based 3-D camera motion estimation with application to video coding," in *Motion Analysis and Image Sequence Processing* (M. I. Sezan and R. L. Lagendijk, ed.), ch. 4, Kluwer Academic Publishers, 1993.

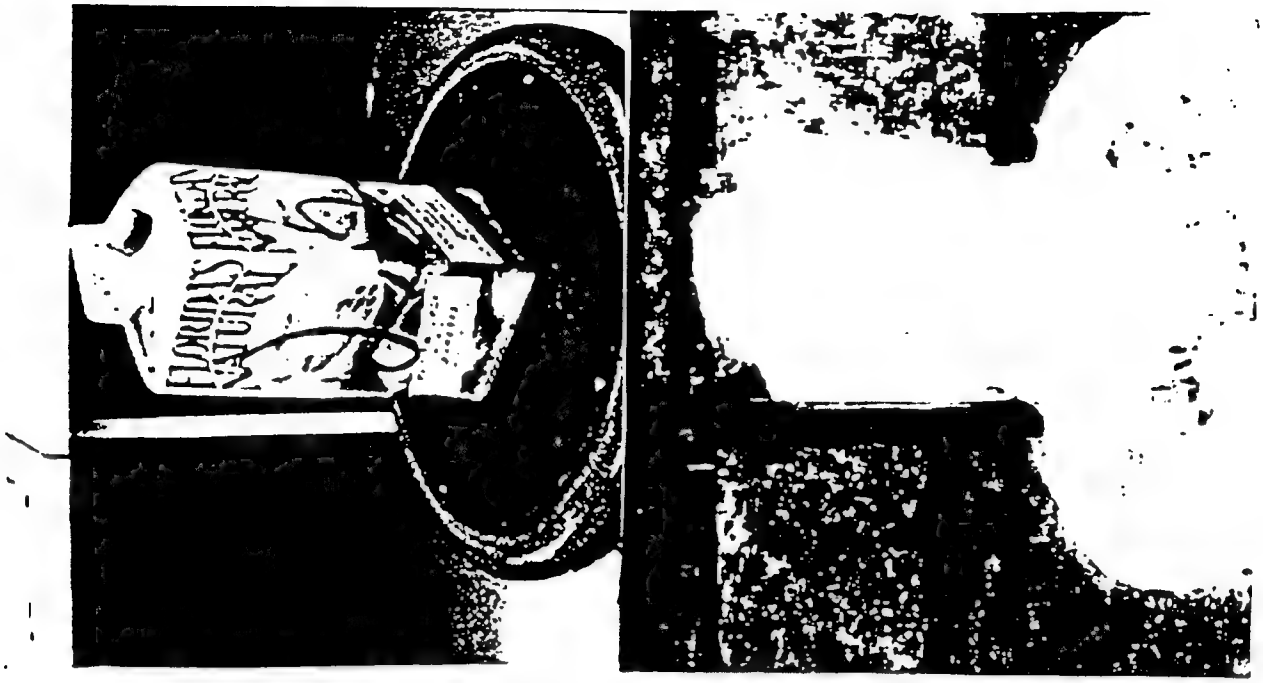
[2] A. Zakhor and F. Lari, "3-D camera motion estimation with applications to video compression and 3-D scene reconstruction," in *Proceedings of the SPIE: Image and Video Processing*, vol. 1903, San Jose, CA, 3–4 Feb. 1993.

[3] R. Y. Tsai and T. S. Huang, "Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI–6, no. 1, pp. 13–26, Jan. 1984.

[4] B. K. P. Horn, *Robot Vision*. Cambridge, MA: MIT Press, 1991.

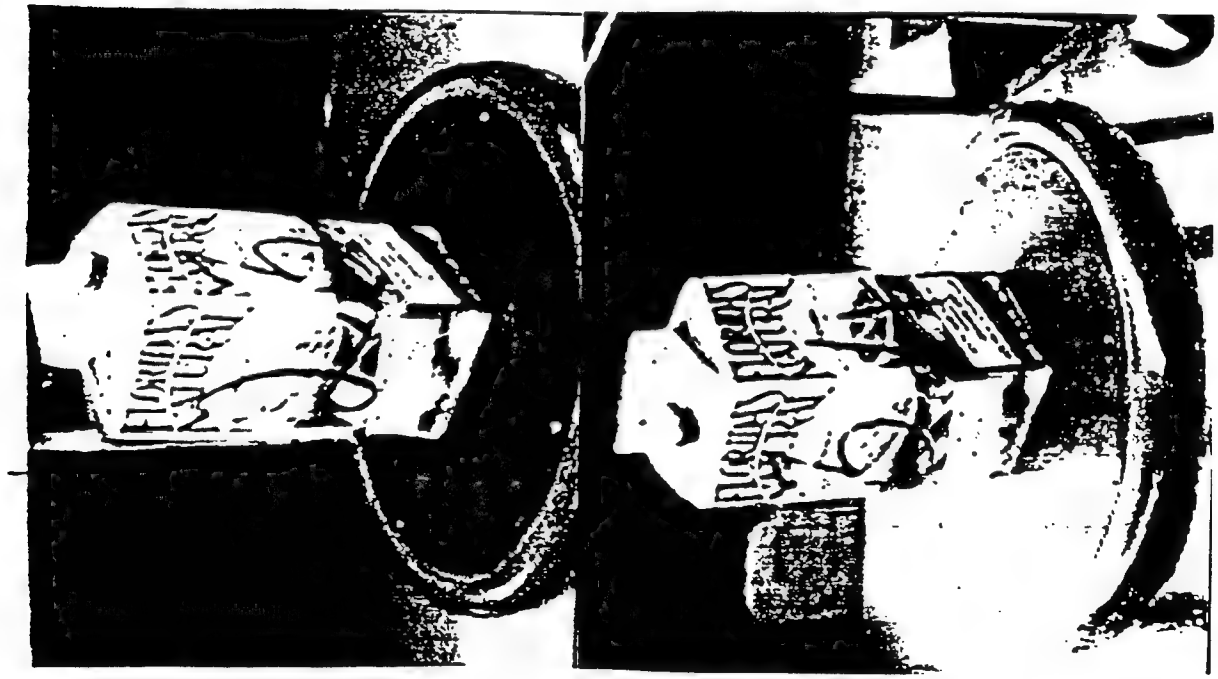
[5] J. Weng, N. Ahuja, T. S. Huang, "Motion and structure from point correspondences with error estimation: Planar surfaces," *IEEE Trans. Sig. Proc.*, vol. 39, no. 12, pp. 2691–2717, Dec. 1991.

[6] P. Anandan, J. R. Bergen, K. J. Hanna, and R. Hingorani, "Hierarchical model-based motion estimation," in *Motion Analysis and Image Sequence Processing* (M. I. Sezan and R. L. Lagendijk, ed.), ch. 1, Kluwer Academic Publishers, 1993.



(a) (b)

Figure 2: *Frame #8 of "Orange Juice" sequence at view 2: (a) intensity, (b) depth map*



(a) (b)

Figure 3: *Typical reconstructed frames along the trajectories. (a) 40% between 8 and 20 (view 2); (b) 40% between 8 and 20 (view 3).*

To Appear in
IEEE Transactions on Circuits and Systems. Part II: Analog and Digital Signal Processing.
Vol. 40, No. 10

An IC Chip of Chua's Circuit *

José M. Cruz and Leon O. Chua†

Abstract — **This paper reports a working microelectronic chip implementation of Chua's circuit. This chip has been designed and fabricated using a 2 μm CMOS technology, with the circuit itself occupying a silicon area of $2.5mm \times 2.8mm$. The chip needs to be powered with a single 9-V battery, is autonomous, and generates the three state variables of Chua's circuit. The proper operation of this chip has been confirmed by experimental reproduction of bifurcation and chaotic phenomena. This microelectronic design of Chua's circuit can be employed as a basic component in the VLSI synthesis of complex circuits making use of chaotic signals, including a class of cellular neural networks and secure communication systems.**

1 Introduction

Electronic circuits exhibiting well-understood bifurcation and chaotic behavior can be exploited as basic components of emerging classes of complex dynamic electronic networks and systems, including cellular neural networks exhibiting spatially chaotic dynamics and secure communication systems based on chaos synchronization.

Chua's circuit [1]–[9] is the simplest autonomous circuit which can exhibit bifurcation and chaos. It has been studied extensively and is one the very few circuits in which a formal proof of the existence of chaos has been accomplished [5]. Moreover, the theoretical and simulated behavior of this circuit can be accurately reproduced experimentally. These factors have made of Chua's circuit a tool for studying and generating chaos, and is being used as a building block for developing other more complex circuits exploiting chaotic and bifurcation phenomena [10] [11].

Several physical implementations of the circuit have been proposed since 1985 [6] [7] [8]. They use discrete components to implement the linear elements and a combination

*This work is sponsored by the Joint Services Electronics Program under contract F49620-93-C-0014.

†The authors are with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720, USA.

of op amps, resistors, diodes, or discrete bipolar transistors to implement the nonlinear element (Chua's diode). Recently, monolithic CMOS implementations of the Chua's diode [2] and the Chua's circuit [1] have been fabricated.

In this paper we report the experimental results and the implementation details of a microelectronic chip implementing Chua's circuit. The linear resistor R is the only element implemented externally, by a potentiometer, to allow the setting of a bifurcation parameter. This chip has been designed and fabricated using a 2 μm double-metal double-poly CMOS technology [15]. The three linear storage elements are implemented with double-poly capacitors, one of them used to emulate the inductor [13]. The resonant frequency of the active LC circuit is approximately 160 KHz. This 8-pin, autonomous chip is powered by a single 9-V bias battery, and generates three output signals representing the state variables of Chua's circuit

The outline of this paper is as follows. Section II gives the chip electrical specifications and its experimental performance. It shows the three projections of the experimental double-scroll Chua's attractor and the experimental bifurcation sequences obtained by modifying two independent parameters. Section III gives the internal structure of the chip and detail the design procedure for a CMOS technology. Section IV presents a numerical simulation of the chip. Finally, Section V gives some concluding remarks and applications of the new chip.

2 Chip Specifications and Experimental Performance

2.1 Parameters of Chua's circuit chip

Chua's circuit, shown in Figure 1, is a third-order circuit. The three variables are the voltages across capacitors C_1 and C_2 and the current through L. They are denoted as v_{C_1}, v_{C_2} and i_L, respectively; and their dynamics are given by:

$$\begin{aligned} C_1\frac{dv_{C_1}}{dt} &= \frac{1}{R}(v_{C_2}-v_{C_1})-g(v_{C_1}) \\ C_2\frac{dv_{C_2}}{dt} &= \frac{1}{R}(v_{C_1}-v_{C_2})+i_L \\ L\frac{di_L}{dt} &= -v_{C_2} \end{aligned} \tag{1}$$

where $g(v_{C_1})$ is the function given in Figure 1(b). Inside the range $(-E_2,E_2)$ of v_{C_1}, in which the circuit normally operates, this function is given by

$$g(v_{C_1}) = m_1 v_{C_1} + \frac{1}{2}(m_2-m_1)(|v_{C_1}+E_1|-|v_{C_1}-E_1|) \tag{2}$$

A particular Chua's circuit is characterized by seven parameters denoted as (C_1, C_2, L, R, E_1, m_1, m_2). They represent, respectively, the values of the two linear capacitors, C_1 and C_2, the linear inductor, L, the linear resistance, R, and finally the first breakpoint, E_1, and the inner slopes, m_1 and m_2, of the Chua's diode driving-point characteristic.

The IC chip reported in this paper implements a Chua's circuit with the seven parameter values given in Table I. Five of these parameters have fixed values (C_1, C_2, E_1, m_1 and m_2), and the two others (L and R) are variable. This allows us to implement with our chip a two-dimensional parameter space of possible Chua's circuits.

Table I. Parameters of the chip

Parameter	Value	Unit
C_1	150	pF
C_2	2000	pF
L	0.33	mH
R	1750	Ω
m_1	-0.41	mA/V
m_2	-0.78	mA/V
E_1	0.7	V

It is shown in [5] how by proper normalization of the three state variables, V_{C_1}, V_{C_2} and I_L, and of the time scale, the set of Chua's circuits with *different dynamics* can be specified with only four parameters (α, β, a, b), instead of seven. However, the control of R and L values still give us access to a two-dimensional parameter space of Chua's circuits. In particular varying R leads to the variation of a combination of β, a and b, while varying L leads to the independent variation of β.

2.2 External description of the chip

Figure 2 shows a photograph of the IC chip of Chua's circuit. The package is an 8-pin DIP, 0.3 inches wide and 0.1 inches interlead. It can be plugged into standard breadboards or op amp sockets. The output pins of the chip are defined in Table II.

Table II. Output pins of the chip

pin no.	*name*	description
1	v_1	output signal v_{C_1}
2	v_2	output signal v_{C_2}
3	v_3	output signal $r_d I_L$
4	v_+	positive terminal for bias
5	v_-	negative terminal for bias
6	$control_1$	terminal for L tuning (optional use)
7	$control_2$	terminal for L tuning (optional use)
8	v_{GND}	output ground reference

The chip is autonomous, and therefore does not require any input signal. The bias is provided by a single 9-V battery connected between terminals v_+ and v_-. To set the two independent bifurcation parameters we use potentiometers R and R_β connected as shown in Figure 3.[1]

The chip generates three output signals representing the three state variables of Chua's circuit. For convenience, the three signals are provided as three voltages, v_1, v_2 and v_3, referenced to a common ground, v_{GND}. The outputs v_1 and v_2 are the voltage across capacitors C_1 and C_2 in Volts. The output v_3 is a voltage proportional to the current though the inductor [2] , according to

$$v_3 = r_d I_L \tag{3}$$

The nominal value of the proportionality constant, r_d, is $-500\frac{V}{A}$.

2.3 Experimental results

Using the above IC chip of Chua's circuit, powered by a 9V-battery, we have generated chaos and bifurcation phenomena.

[1]The potentiometer R_β is used only as a convenient means of changing the bifurcation parameter β [5] in the experimental results presented in this paper. However, this potentiometer R_β is *not* necessary for the operation of the chip. An alternative way to change β, if desired, is by applying directly a voltage bias reference to pin $control_1$, which controls the value of the internal voltage-controlled-inductor. This later approach, of using electronic control of the internal L, may be more convenient for those using this chip as part of a larger electronic system.

[2]In previous experimental implementations of Chua's circuit this third signal, representing the current though the inductor, has been generated by measuring the voltage drop in a small resistor connected in series with the inductor, an approach which may distort the dynamic behavior of the circuit. In our implementation, as we will demonstrate in Section III, the signal v_3 is generated without introducing any artifact affecting the dynamics of Chua's circuit.

2.3.1 Chaos generator

Using the nominal values of $R = 1750\Omega$ and $R_\beta = 20K\Omega$ the chip works as a generator of chaotic signals. The chip operates in the double-scroll region. Figure 4 shows the experimental time waveforms of the three output state variables v_1, v_2 and v_3. Figure 5 shows the three experimental Lissajous figures. They represent the projection of the chaotic strange attractor onto the (v_1, v_3), (v_2, v_1), and (v_2, v_3) planes.

Figure 6(a) shows a photograph with a magnified detail of the central part of the (v_1, v_3) projection of the double-scroll Chua's attractor. It corresponds to the same conditions of the lower left photograph of the previous figure. Figure 6(b) shows a further magnification of the region. It is possible to distinguish some of the individual trajectories that thickly fill the outer surface of the attractor during the 1/8s of the time-exposure photograph. In the center of this photograph, a perspective of the trajectories going along the inner part of the spiral cylinder can also be observed.

Note that for the same parameters there is another possible solution, that is a large limit cycle associated with the outer slopes of the nonlinear element. However, this solution can only be observed if the initial conditions for the state variables is forced to be far away from the origin, outside the basin of attraction of the double-scroll Chua's attractor. Normally, when the chip is powered, the state variables are inside the basin of attraction of the chaotic attractor, because, due to current leakage, the storage capacitors are initially discharged. Figure 7 shows a photograph superimposing the two solutions.

2.3.2 R Bifurcation sequence

The well-known bifurcation sequence obtained by decreasing the resistor value R has been experimentally reproduced. As an example, Figure 8, shows the experimental Lissajous figures in the (v_1, v_3) plane. R_β is kept constant at the nominal value of $20K\Omega$. As R is decreased from 2050Ω to 1568Ω we observe periodic behavior emerging from a stable equilibrium point, then a period-doubling sequence, a spiral Chua's attractor and finally a double-scroll Chua's attractor. As R is decreased further the double-scroll Chua's attractor shrinks in size and its central region becames thinner. As R decreases below 1568Ω, the double-scroll Chua's attractor and the saddle-type periodic orbit collide with each other. At that point the only solution is the large limit cycle determined by the outer segments of the Chua's diode characteristic.

2.3.3 β Bifurcation sequence

This chip also allows independent change of the bifurcation parameter β [5]. Using the configuration of figure 3, the parameter β can be increased by decreasing the value of the potentiometer R_β. As an example, Figure 9, shows the experimental Lissajous figures in the (v_2, v_1) plane. R is kept constant at the nominal value of 1750Ω. As R_β is decreased from $26K\Omega$ to $18K\Omega$ we observe as before periodic behavior emerging from a stable equilibrium point, then a period-doubling sequence, a spiral Chua's attractor and finally a double-scroll Chua's attractor. Observe that now the chaotic attractor does not decrease in size as we change the bifurcation parameter.

3 Internal structure of the chip

The chip described in the previous section has been implemented using a CMOS process. In this section, we present the internal structure of the chip, and describe the design procedure used. This design procedure can be used for the VLSI implementation of Chua's circuit in a different technology or with different parameters. This section is structured as follows. In the first part we give the internal structure of the chip at the network element level. Then, we detail the design of ecah of these elements at the transistor level for a CMOS technology. Finally, we present the physical structure of the entire chip.

3.1 Internal Network level structure

Figure 10 gives the network schematic of the Chua's circuit implemented in the chip. It contains the nonlinear 2-terminal Chua's diode N_R, three capacitors C_1, C_2 and C_3, and and a gyrator G with admittance matrix given by

$$Y_G = \begin{bmatrix} 0 & g_d \\ -g_c & 0 \end{bmatrix} \tag{4}$$

The gyrator terminated at its right port by capacitor C_3 looks like an inductor of value

$$L = \frac{1}{g_c g_d} C_3 \tag{5}$$

at its left port.

This circuit has 3 nodes. The voltage at these nodes, denoted by v_1, v_2 and v_3 are available at the chip output. In spite of the fact that we have introduced an extra node, with respect to the circuit of Figure 1, we have *not* introduced any other state

variable into the circuit. The two new circuit variables introduced in the new circuit, v_3 and i_{C_3} have values determined by

$$\begin{aligned} v_3 &= -\frac{i_L}{g_4} \\ i_{C_3} &= g_c v_2 \end{aligned} \tag{6}$$

Therefore, the circuits in Figure 1 and 10 are equivalent. The latter is more suitable for VLSI implementation. Besides, the availability of the third state variable as a voltage allows us to experimentally measure this variable without introducing any measuring circuitry that could modify the dynamics of Chua's circuit.

The nonlinear resistor and the gyrator are active network elements and therefore need to be biased. We use a bias scheme in which only one external battery is used. Figure 11 shows the entire circuit including the bias. The floating external battery is connected to terminals v_+(pin 4) and v_-(pin 5). These terminals are internally connected to the positive and negative supply of the Chua's diode, of the gyrator, and of the bias circuit shown at the right of the figure. This bias circuit generates, at its low resistance output, a signal v_{GND} (pin 8), with voltage value in the middle of the values at the v_+ and v_- nodes. This voltage v_{GND} is considered to be our ground.

3.2 CMOS implementation

Figure 12 shows the architecture of the entire circuit using CMOS amplifiers and capacitors. The two operational transconductance amplifiers A and B, in positive feedback configuration, implement the Chua's diode [2]; the two back-to-back transconductance amplifiers C and D implement the gyrator [14]; and the serial set of CMOS diodes and the operational amplifier in negative feedback configuration implement the bias circuit.

For the implementation we have used a $2\mu m$ double-metal double-poly CMOS technology. The most relevant parameters of this technology are given in Table III. The capacitors have been implemented directly by using the two poly layers as capacitor plates (capacitance per unit area is $470pF/mm^2$ in our technology). The voltage-mode operational amplifier has been designed using the two stage miller-compensated topology [12]. The four operational transconductance amplifiers have been designed using a topology based on simple differential pairs, as this gives the maximum effective frequency response and minimal input noise.

Table III. Technological Data

Parameter	N-channel	P-channel	Unit
V_{th}	1.0	0.8	V
μC_{ox}	47	23	$\mu A/V^2$
γ	1.06	0.45	$\sqrt{V}$
ΔL	0.54	0.42	μm
ΔW	0.07	0.17	μm

Figure 13 shows the transistor schematic topology used for the OTAs. This topology is the same for all the OTAs, but each is designed to obtain different transconductance characteristics. The transconductance gain of each of them, at the origin, is denoted as g_a, g_b, g_c and g_d, respectively. Table IV gives their nominal values.

The transconductance amplifiers A and B implement the nonlinear Chua's diode. They determine the parameters m_0, m_1 and E_1. The transconductance g_a and g_b of OTA A and OTA B should be equal to $m_1 - m_2$ and $-m_1$, respectively. The output current of OTA A is limited to a constant value when v_1 reach the breakpoint $E_1 = 0.7$ V. In this chip the OTA A and B have fixed characteristics, that are set by the self-bias circuit at the left of Figure 13. [3]

The necessary nonlinearity of the circuit is produced by the cut-off of just a pair of transistors of the differential pair of OTA A: T101 (for $v1 \leq E_1$) or T102 (for $v1 \geq E_1$). As transitions from cut off to conduction can cause small delays, we want to prevent any other transistors in the signal path from cutting off. We achieve this by shorting the drains of T101 or T102 of OTA A with the equivalent transistors of OTA B. This connection (which does not appear in Figure 12 to avoid clutter) is equivalent to the parallel connection of the current mirror of both OTAs [4]. The current bias of OTA B will always maintain all current mirrors in conduction. Using this scheme we can obtain a driving point characteristic for the Chua's diode which does not show any measurable hysteresis phenomena at the frequency of operation centered in the 160 KHz range. The design procedure to determine all the transistor dimensions of these two amplifiers can be found in [2]. For our particular bias levels the final dimensions values are given in Table V.

The transconductance amplifiers C and D implement the gyrator. They determine

[3] They can be adjusted if an external pin is assigned to the control line of OTA A. We have recently used that scheme in a Chua's diode chip prototype, and we have successfully used it to experimentally reproduce bifurcation phenomena by continuously varying the slope m_0 (bifurcation parameter a).

[4] Equivalently, is also possible to merge all the current mirrors of OTA A with those of OTA B, but increasing accordingly the width of their transistors

the parameters L according to equation (5). The capacitor C_3 has a fixed value of 269 pF. The transconductances g_c and g_d are controllable in order to get a variable inductor. Their nominal values, given in Table IV, are obtained when the control line of OTA C (pin 7 of the chip) is left open, and the control line of OTA D (pin 6 of the chip) is connected to $R_\beta = 20K\Omega$ as indicated in Figure 3. Under this condition the gyration ratio is 1.23×10^6 and the emulated inductor has value of 0.33mH. This value can be changed in a ± 50 range by R_β adjustment. All transistor dimensions of these amplifiers are shown in Table V.

The LC circuit formed by C_2 and the emulated inductor has a quality factor of $Q = 78.5$, which is actually higher that what is usually obtained by using discrete components[5]. This has been achieved by designing gyrator amplifiers with very large ratios between their transconductances and their output conductances.

Table IV. Transconductance Values

OTA	transconductance	Unit
A	0.37	mA/V
B	0.41	mA/V
C	0.41	mA/V
D	2.00	mA/V

[5] A typical series resistance of 13Ω in the inductor will degrade the quality factor to approximately $Q = 25$

Table V. Mask dimensions of the internal transistors of the OTAs

Device	W (μm)			L (μm)
	OTA A	OTA B and C	OTA D	
$T_{101A}, T_{101B}, T_{102A}, T_{102B}$	40	15	15	6
$T_{103A}, T_{105A}, T_{106A}$	280	400	400	4
$T_{103B}, T_{105B}, T_{106B}$	280	400	400	2
T_{104A}	280	400	3200	4
T_{103B}	280	400	3200	2
T_{107A}	140	200	200	4
T_{107B}	140	200	200	2
T_{108A}	138	200	1600	4
T_{108B}	138	200	1600	2
T_{109A}	100	476	476	6
T_{301A}	50	50	50	10
T_{301B}	100	100	100	10
T_{302A}, T_{302B}	33	33	33	10

An conventional operational amplifier and a set of diodes are used to implement the bias circuit. The diodes are made by gate-to-drain connected transistors. They are sized so that their midpoint voltage (v_{GND}) is just in the middle of the values at the v_+ and v_- nodes. The accuracy of this voltage division is not critical, as the circuit is, to first order, insensitive to variations in the DC voltage difference between v_{GND} and the supply nodes. AC variations are minimized to less than 6 mV by the use of a high gain conventional operational amplifier with negative feedback.

3.3 Physical structure

The circuit has been fabricated in 2 μm CMOS technology of ORBIT Semiconductors [15]. Figure 14 shows a micrograph of the fabricated circuit. It occupies a silicon area of $2.5mm \times 2.8mm$. All the active circuitry is the central part of the upper side of the die. The three rectangular blocks at the bottom, from left to right, are respectively capacitors C_1, C_2 and C_3.

4 Numerical simulations

The experimental results are validated with numerical simulations. As as an example, figure 15 shows a device-level numerical simulation of the chip in chaotic operation,

with the same conditions used to obtain the expermental results of Figure 5. The correspondence with the experimental data shown earlier is excellent.

5 Concluding Remarks

In this paper we have presented a working microelectronic Chua's circuit which produces chaotic signals whose experimental dynamics are in in close concordance with theoretical and numerical predictions based on the ideal Chua's equation [5]. The circuit occupies an area of 7 square millimeters in $2\mu m$ CMOS technology. In a large die it is possible to place 57 of our circuits. The number of possible circuits on a chip can be increased to about 600 by applying the scaling rules given in Appendix I.

Our major motivation for this work was the need of microelectronic circuits that could be used as a building block of several classes of systems that require the use of chaotic behavior. Some examples are secure communication systems based on chaos synchronization [10], and network arrays [11] [16] with spatially chaotic dynamics. The successful implementation of these systems relies upon the availability of a chaotic electronic component exhibiting experimental dynamics that closely resembles a mathematical model and that can be accurately controlled. We hope that our design will facilitate the VLSI implementation of these emerging classes of circuits and systems making use of chaotic phenomena.

Acknowledgment

The authors would like to thank Dr. M. P. Kennedy for useful discussions on the bias circuit, and to C. W. Wu for photographing the chip.

Appendix I: Scaling rules for high-density VLSI implementation of Chua's circuits.

With our present design it is possible to design chips containing up to 57 circuits (assuming a large die size of $20mm \times 20mm$). Those interested in higher integration densities can scale down the capacitors values and the current levels. The simplest scaling scenario is a linear scaling, in which the new values of capacitors and conductances are kC_1, kC_2, kC_3, km_1, km_2 and $\frac{1}{k}R$, where k is the scaling parameter. The scaling of the capacitors is done simply by reducing the area. The scaling of the conductances is done by reducing proportionally the width of the transistors of the input stage of the OTA A and B. The gyrator design should be unchanged.

After doing this scaling the circuit will operate at the same frequency and we will get the same voltage levels for all variables. All the currents, however, will be scaled according with the same factor k. As both the capacitors and the currents are scaled with the same factor the GBW of all transconductance amplifiers does not degrade in first order, and remain above the operating frequency of the circuit.

The area estimate for the entire microelectronic Chua's circuit in square millimeters it is equal to $0.5 + 6.5(k)$, where $0.5mm^2$ is the area of the nonscaling elements and $6.5mm^2$ is the original area of the scalable elements. The lowest value of k is determined by several factors, including noise degradation, parasitic capacitances that affect the dynamics, and degraded amplifier phase margins and linearity ranges. If we consider a realistic lower limit of $k = 0.02$, this gives 634 Chua's circuits per chip. Higher densities can be achieved by using more advanced technologies.

References

[1] M. Delgado Restituto and A. Rodriguez Vazquez, "A CMOS monolithic Chua's Circuit," *Journal of Circuits, Systems and Computers*, vol. 3, No. 2., June 1993.

[2] J. M. Cruz and L. O. Chua, "A CMOS IC Nonlinear Resistor for Chua's Circuit," *IEEE Transactions on Circuits and Systems*, vol. 39, No. 12, pp. 985–995, December 1992.

[3] L. O. Chua, "The Genesis of Chua's Circuit," *International Journal of Electronics and Communications*, vol. 46, No. 4., pp 250–257, 1992.

[4] T. Matsumoto, "A chaotic attractor from Chua's circuit," *IEEE Transactions on Circuits and Systems*, vol. CAS-31, no. 12, pp. 1055–1058, Dec. 1984.

[5] L. O. Chua, M. Komuro, and T. Matsumoto. "The Double Scroll family, parts I and II," *IEEE Transactions on Circuits and Systems*, vol. CAS-33, no. 11, pp. 1073–1118, Nov. 1986.

[6] G. Q. Zhong and F. Ayrom, "Experimental confirmation of chaos from Chua's circuit," *International Journal of Circuit Theory and Applications*, vol. 13, no. 1, pp. 93–98, Jan. 1985.

[7] G. Q. Zhong and F. Ayrom, "Periodicity and Chaos in Chua's circuit." *IEEE Transactions on Circuits and Systems*, vol. CAS-32, no. 5, pp. 1073–1118, May 1985.

[8] M. P. Kennedy, "Robust op amp realization of Chua's circuit," *Frequenz*, vol. 46, No. 3-4, pp. 66–80, March-April 1992.

[9] R. N. Madan. Guest Editor, "Chua's Circuit: A Paradigm for Chaos," *Journal of Circuits, Systems and Computers.* Part I: vol. 3, No. 1, March 1993. Part II: vol. 3, No. 2., June 1993.

[10] Lj. Kocarev, K. S. Halle, K. Eckert, L. O. Chua and U. Parlitz, "Experimental Demonstration of Secure Communications via Chaotic Synchronization," *International Journal of Bifurcation and Chaos*, vol. 2, No. 3, pp. 709–714, Sep. 1992.

[11] V. Perez-Munuzuri, V. Perez-Villar and L. O. Chua, "Propagation Failure in Linear Arrays of Chua's Circuits," *International Journal of Bifurcation and Chaos*, vol. 2, No. 2, pp. 403–406, June 1992.

[12] P. R. Gray and R. G. Meyer, "MOS Operational Amplifier Design – A Tutorial Overview," *IEEE Journal of Solid-State Circuits*, vol. 17, no. 6, pp. 969–982, Dec. 1982.

[13] Y. T. Wang and A. A. Abidi, "CMOS Active Filter Design at Very High Frequencies," *IEEE Journal of Solid-State Circuits*, vol. 25, no. 6, pp. 1562–1574, Dec. 1990.

[14] R. L. Geiger and E. Sanchez-Sinencio, "Active Filter Design Using Operational Transconductance Amplifier: A tutorial," *IEEE Circuits and Devices Magazine*, vol. 1, no. 2, pp. 20–32, March 1985.

[15] C. Tomovich (editor), "MOSIS User Manual. Release 3.1," *The MOSIS Service*, University of Sourthern California, 1991.

[16] Leon O. Chua and Lin Yang, "Cellular Neural Networks: Theory," *IEEE Transactions on Circuits and Systems*, vol. 35, no. 10, pp. 1257–1272, October 1988.

Figure Captions

Figure 1. (a) Chua's Circuit; (b) Driving-point characteristic of Chua's Diode.

Figure 2. Photograph of the IC Chip.

Figure 3. A typical operating configuration of the chip, with the bias battery, and two external potentiometers to set two bifurcation parameters.

Figure 4. Experimental time-domain waveforms of the state variables. (a) v_1 vs. *time*; (b) v_2 vs. *time*; (c) v_3 vs. *time*. (Vertical scales are 1V/div.; horizontal scale is $50\mu s$/div.)

Figure 5. Experimental Lissajous figures of the double-scroll Chua's attractor. All scales are $500mV$/div.

Figure 6 Detail around the origin of the projection of the double-scroll Chua's attractor onto the (v_1, v_3) plane.

a) Vertical and horizontal scale is $200mV$/div.;

b) Vertical and horizontal scale is $100mV$/div.

Figure 7. Experimental limit cycle outside the double-scroll Chua's attractor. Projection into the (v_1, v_3) plane. (Vertical scale is $2V$/div.; horizontal scale is $2V$/div.)

Figure 8. R-bifurcation sequence. Experimental v_1 vs v_3 Lissajous figures. (Vertical scale is $1V$/div.; horizontal scale is $500mV$/div.).

(a) $R = 2050\Omega$, period one;

(b) $R = 2015\Omega$, period two;

(c) $R = 2009\Omega$, period n;

(d) $R = 1974\Omega$, spiral Chua's attractor;

(e) $R = 1887\Omega$, double-scroll Chua's attractor after birth;

(f) $R = 1568\Omega$, double-scroll Chua's attractor before dying.

Figure 9. β-bifurcation sequence. Experimental v_2 vs v_1 Lissajous figures. (Vertical scale is $500mV$/div.; horizontal scale is $1V$/div.)

(a) $R_\beta = 26.0K\Omega$, period one;

(b) $R_\beta = 25.0K\Omega$, period two;

(c) $R_\beta = 24.6K\Omega$, period four;

(d) $R_\beta = 24.0K\Omega$, spiral Chua's attractor;

(e) $R_\beta = 22.0K\Omega$, double-scroll Chua's attractor after birth;

(f) $R_\beta = 18.0K\Omega$, double-scroll Chua's attractor before dying.

Figure 10. Network schematic of the circuit.

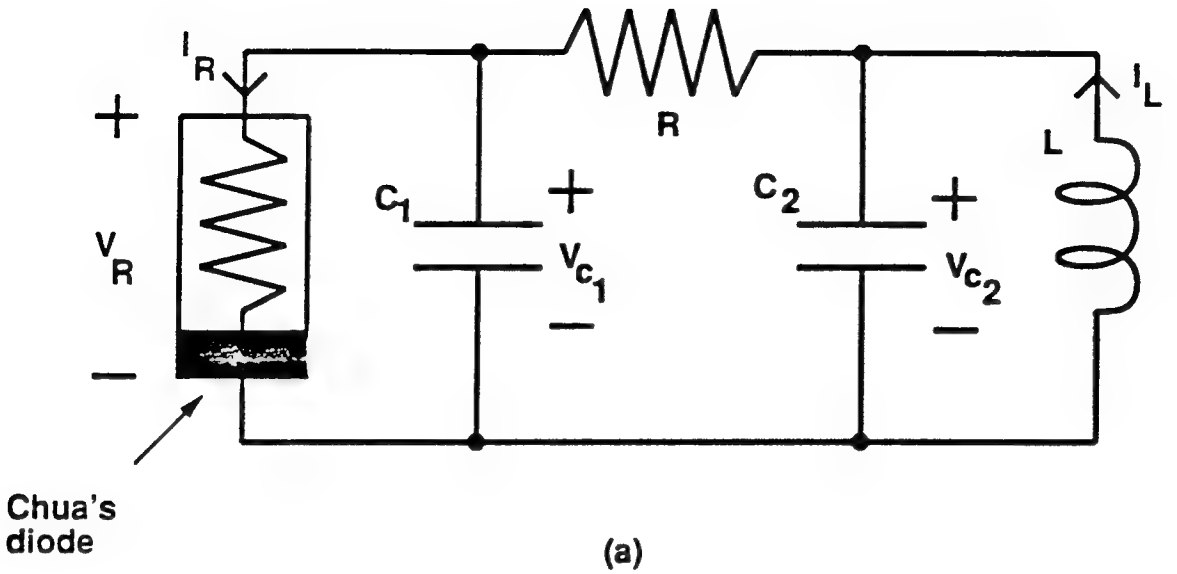
Figure 11. Network schematic of the circuit including bias.

Figure 12. Architecture of the chip.

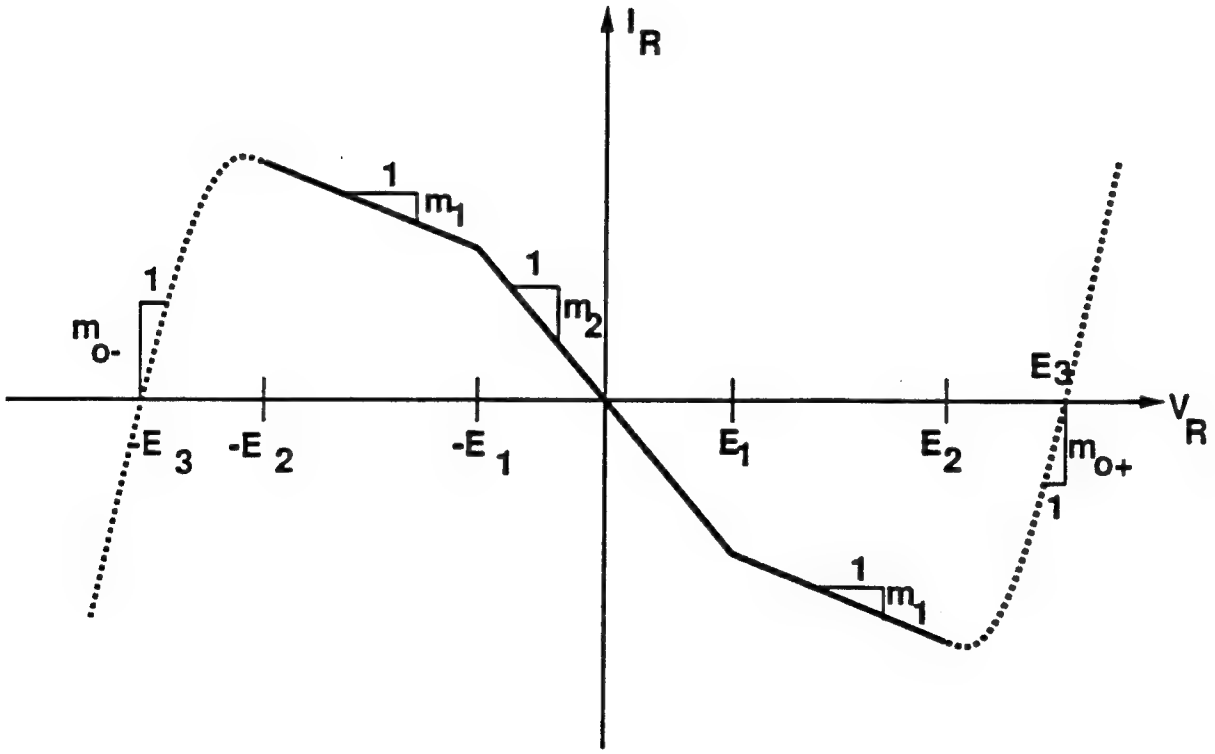
Figure 13. Transistor schematic of the OTAs.

Figure 14. Micrograph of the fabricated circuit.

Figure 15. Lissajous figures of the double-scroll Chua's attractor obtained by electrical simulation. All scales are $500mV$/div.


I R
R
I L
L
C1
C2
V R
Vc1
Vc2
Chua's
diode

(a)


I R
1
m1
1
m2
1
mo-
E3
V R
-E 3
-E 2
-E 1
E1
E2
mo+
1
1
m1

(b)

Figure 1

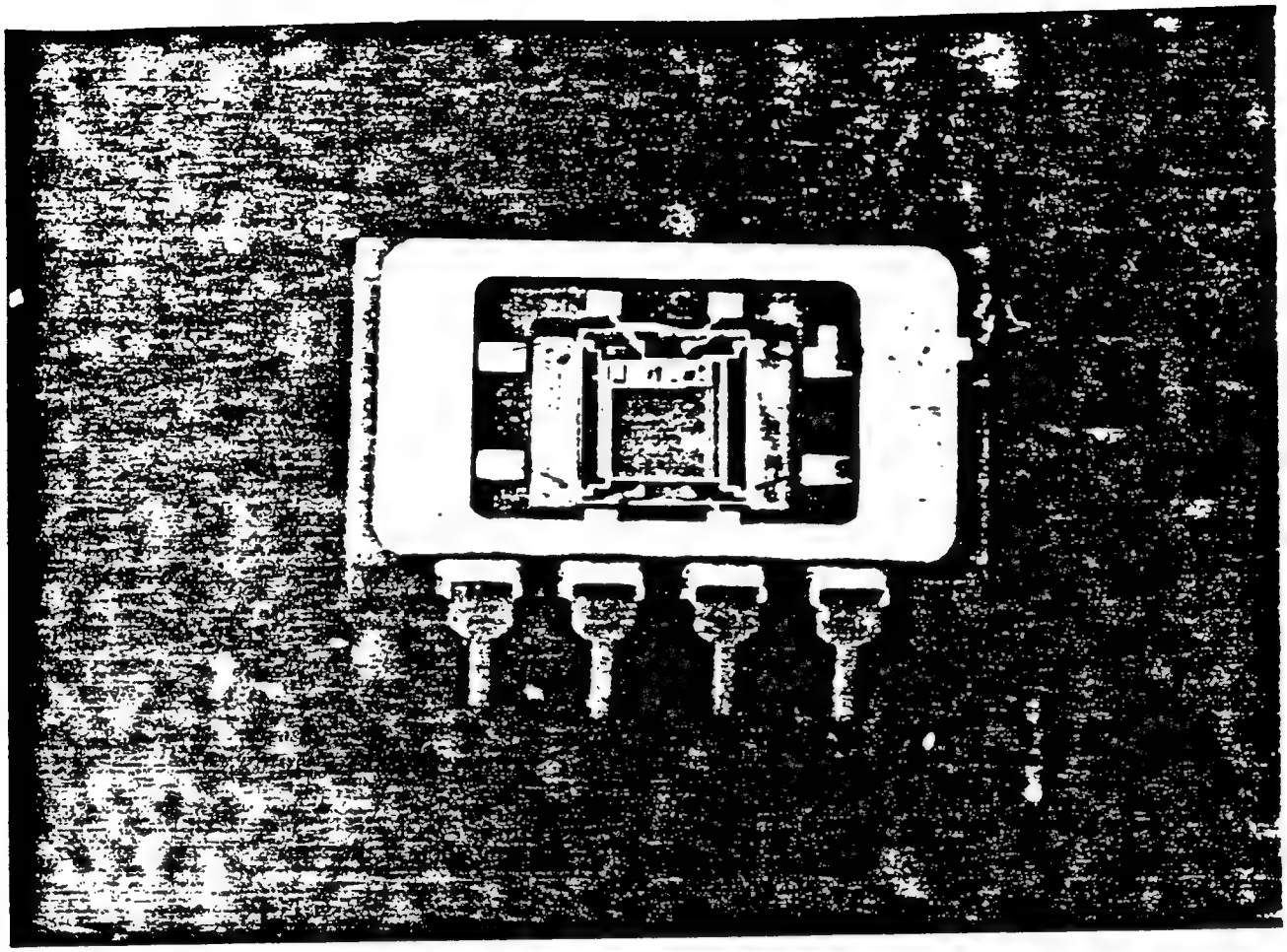


Figure 2

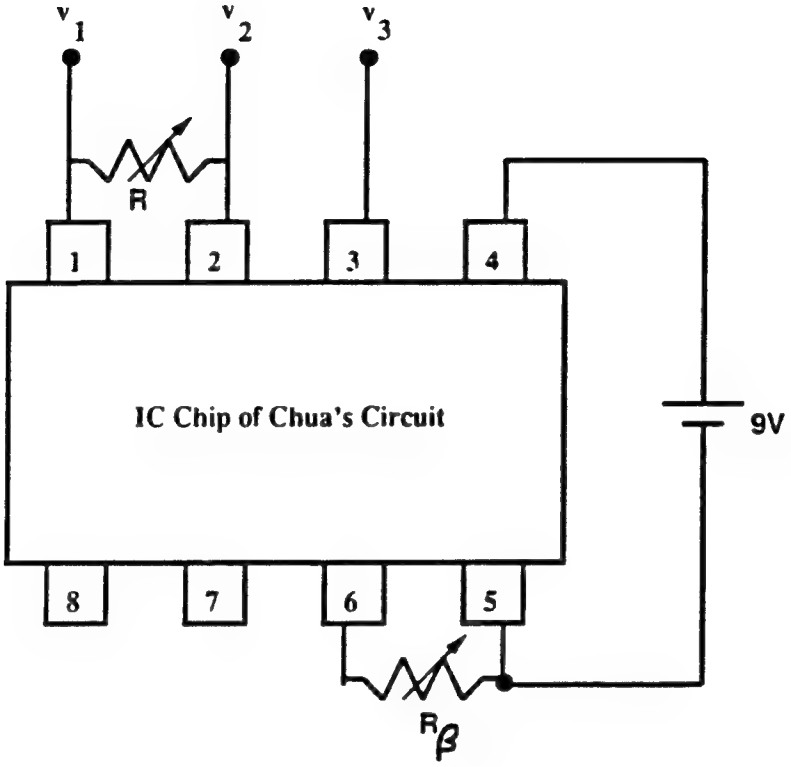
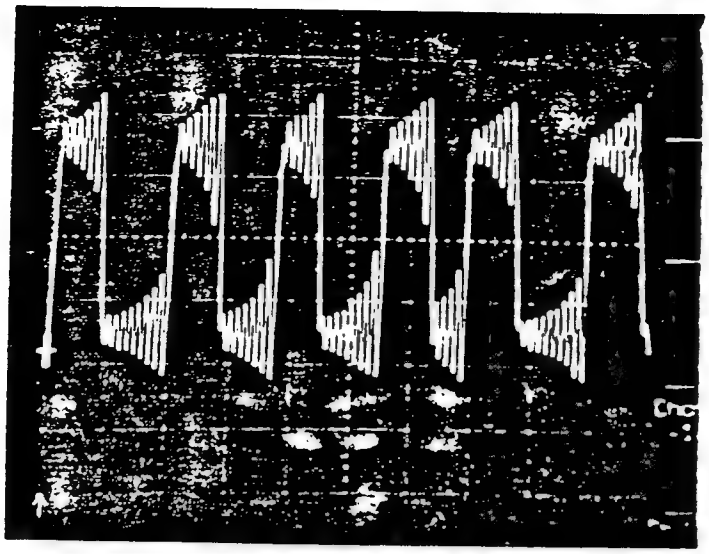
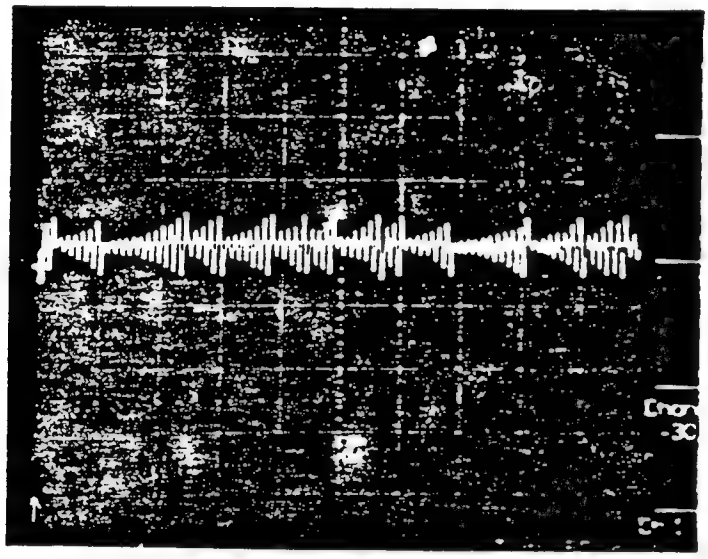


Figure 3

(a)



(b)



(c)

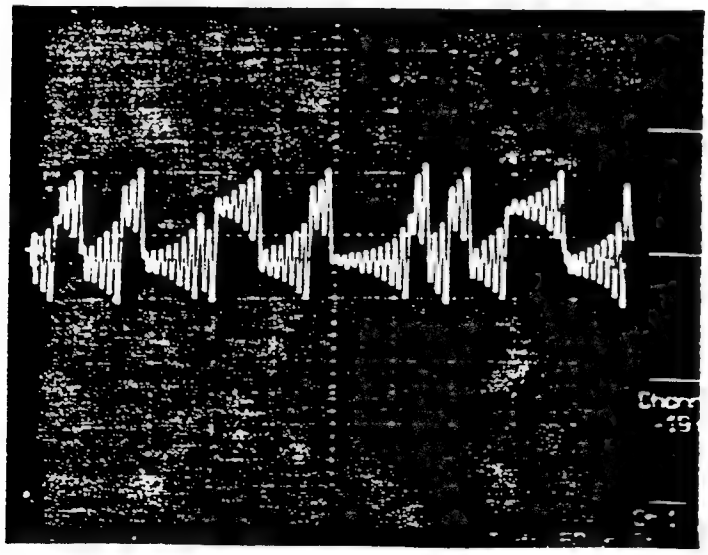


Figure 4

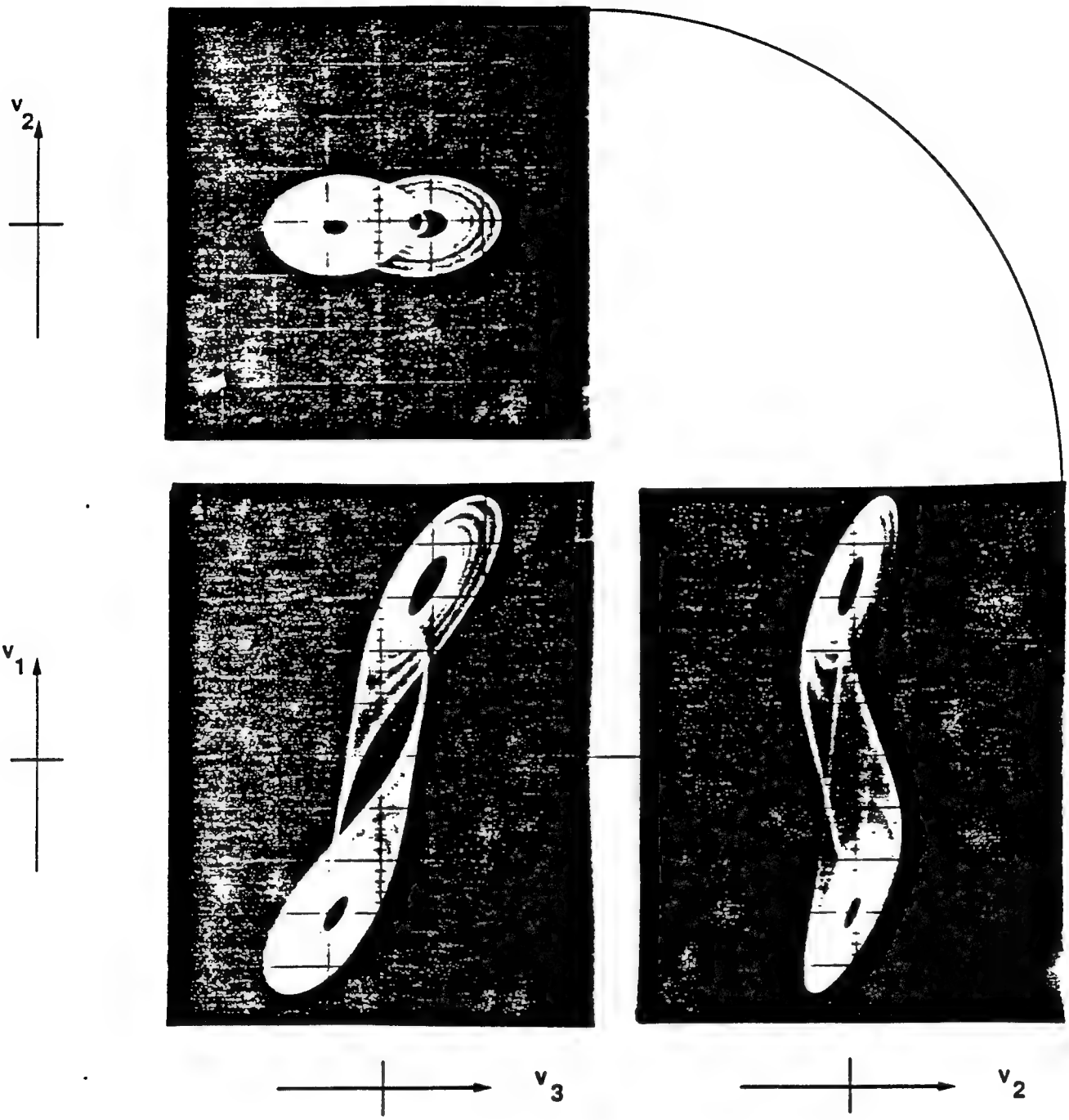
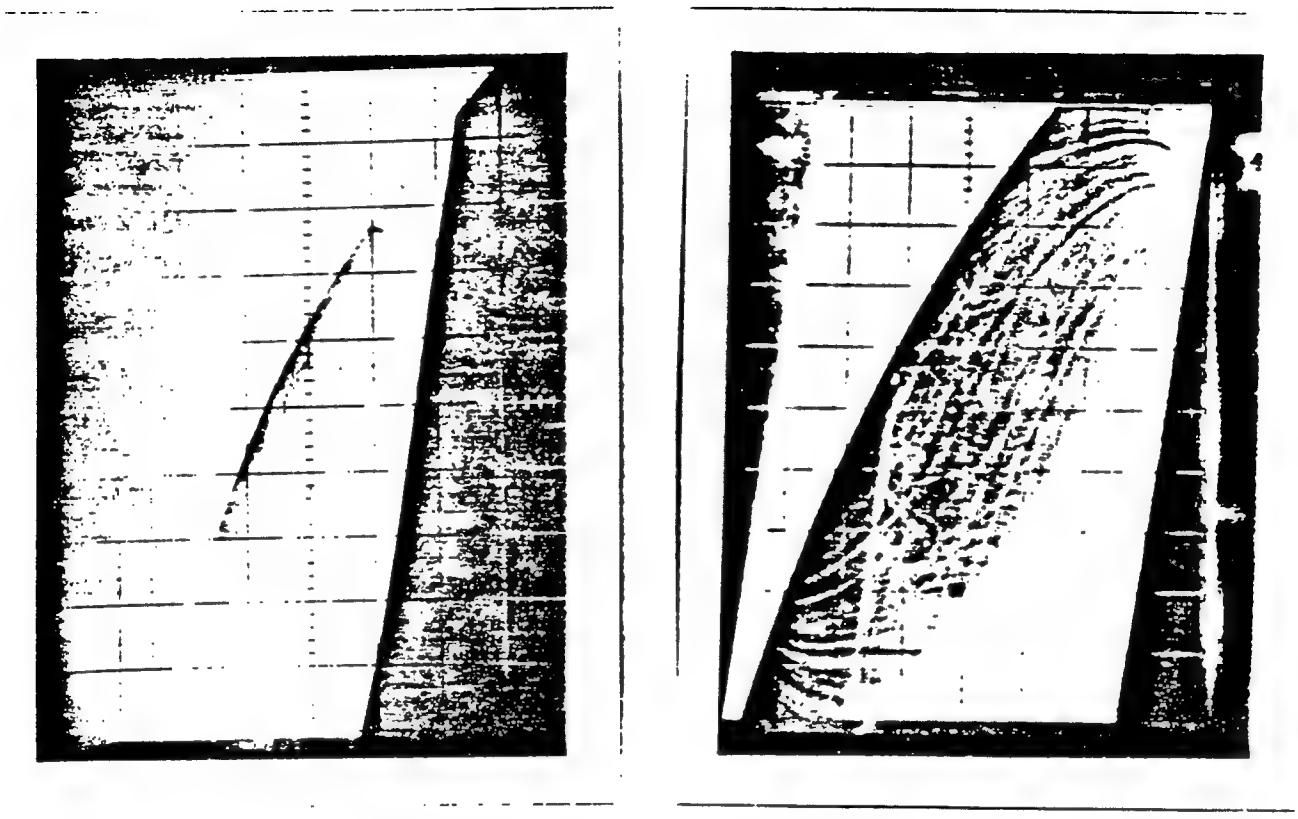


Figure 5



(a) (b)

Figure 6

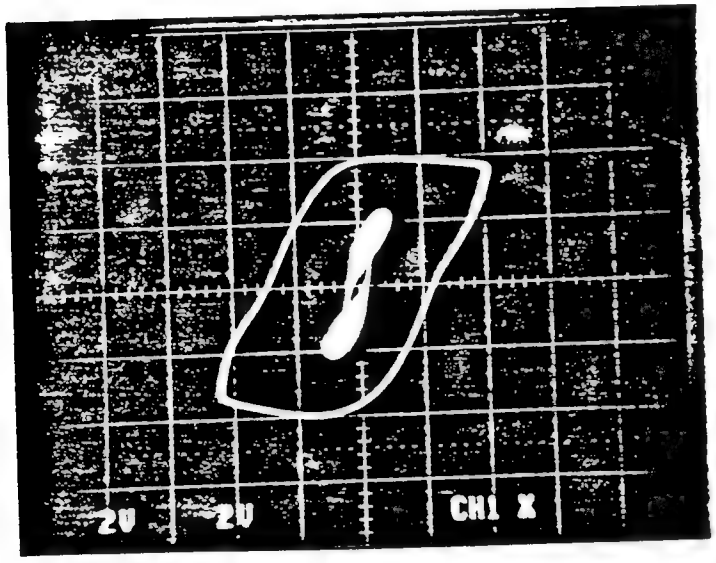
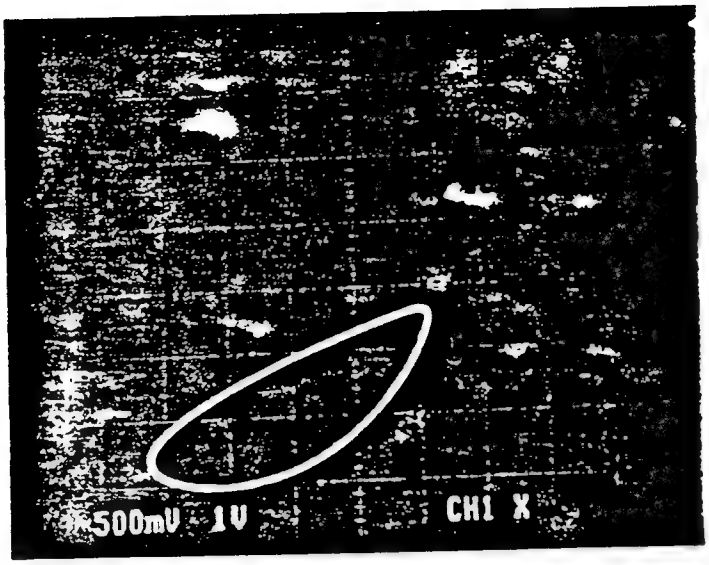
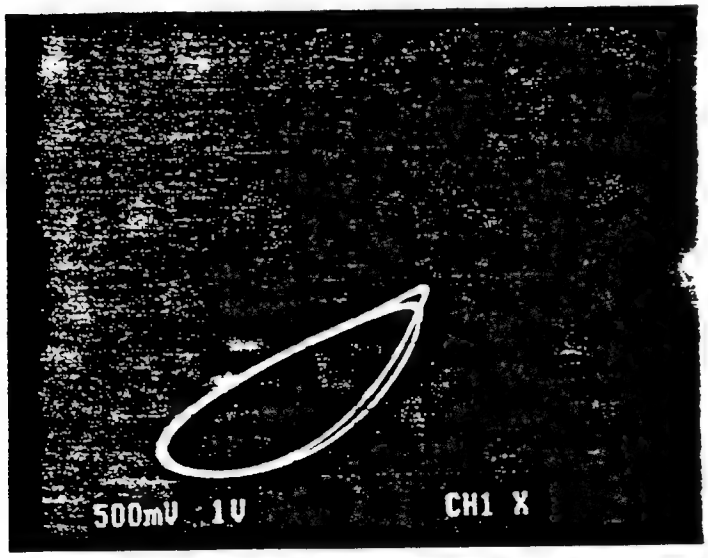


Figure 7

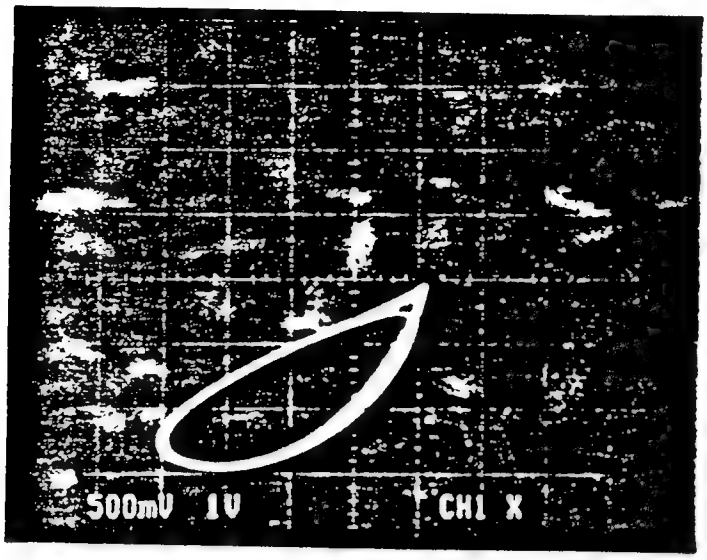


(a)

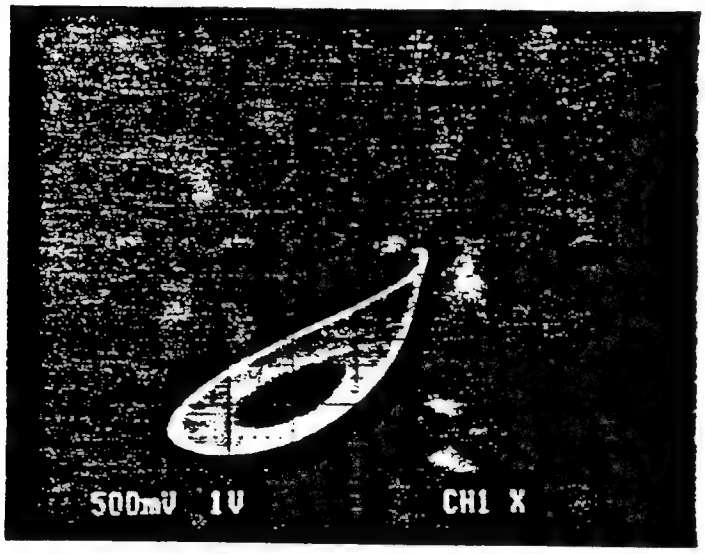


(b)

Figure 8

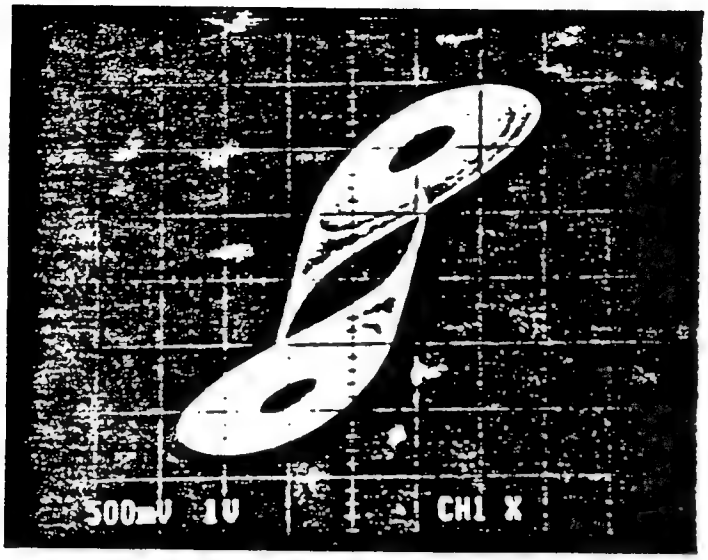


(c)

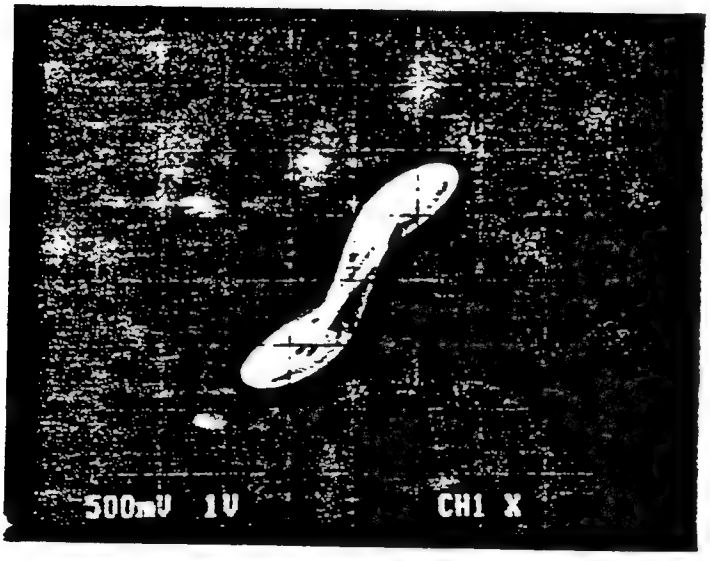


(d)

Figure 8

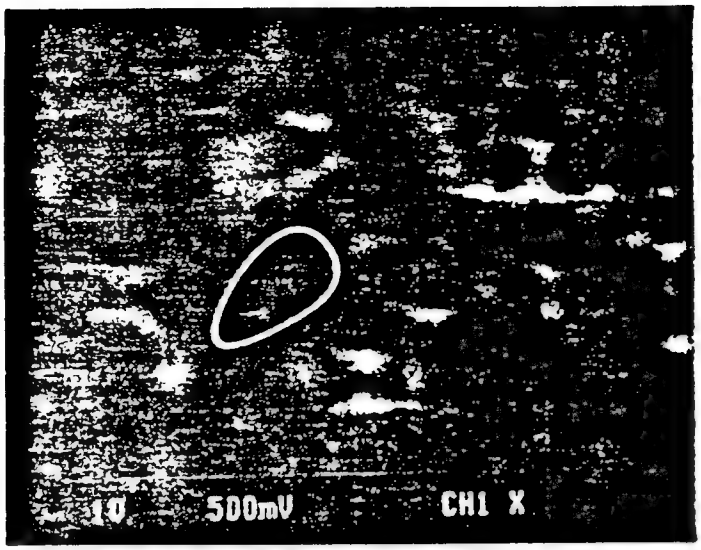


(e)

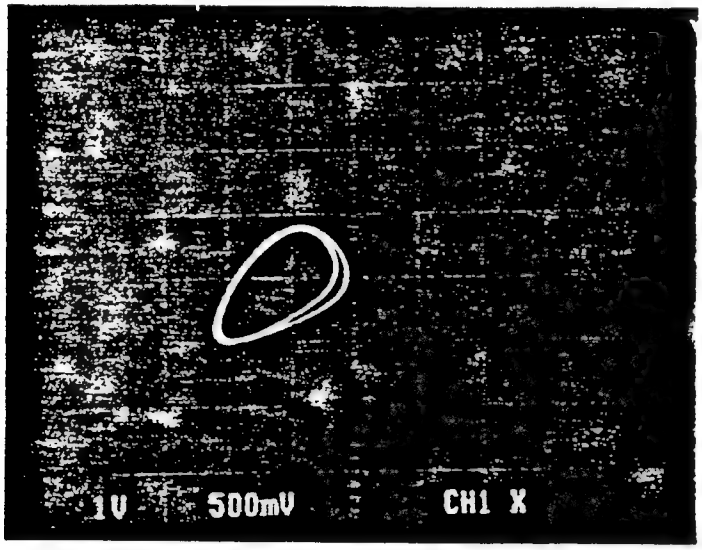


(f)

Figure 8

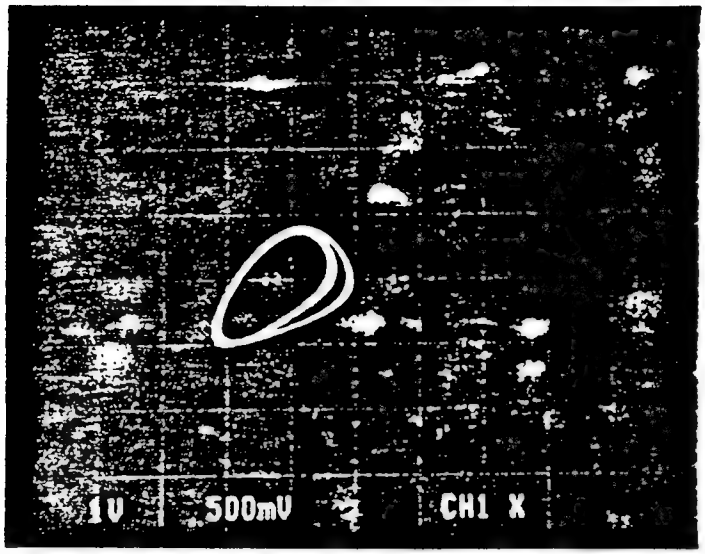


(a)

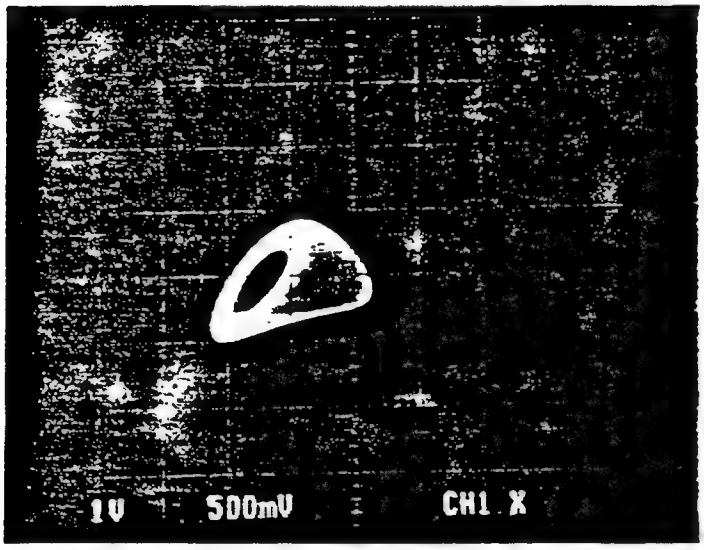


(b)

Figure 9

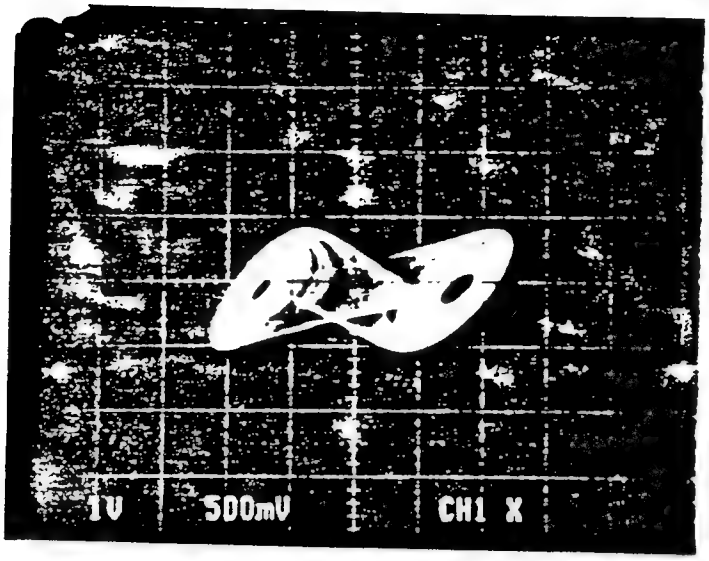


(c)

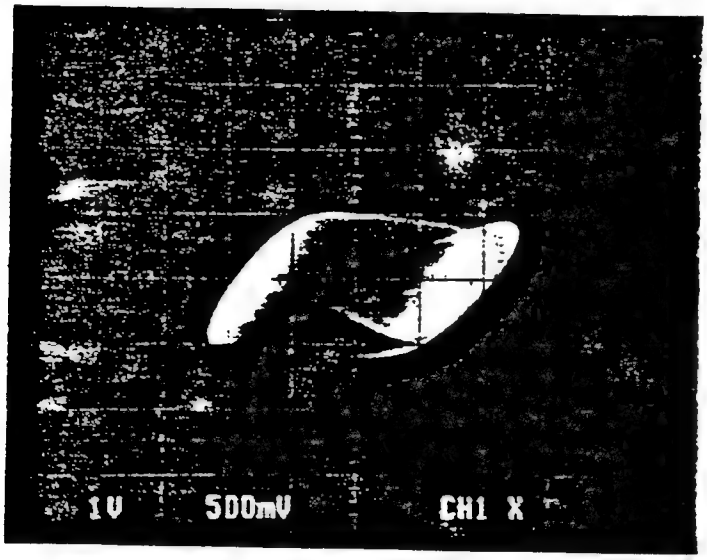


(d)

Figure 9



(e)



(f)

Figure 9

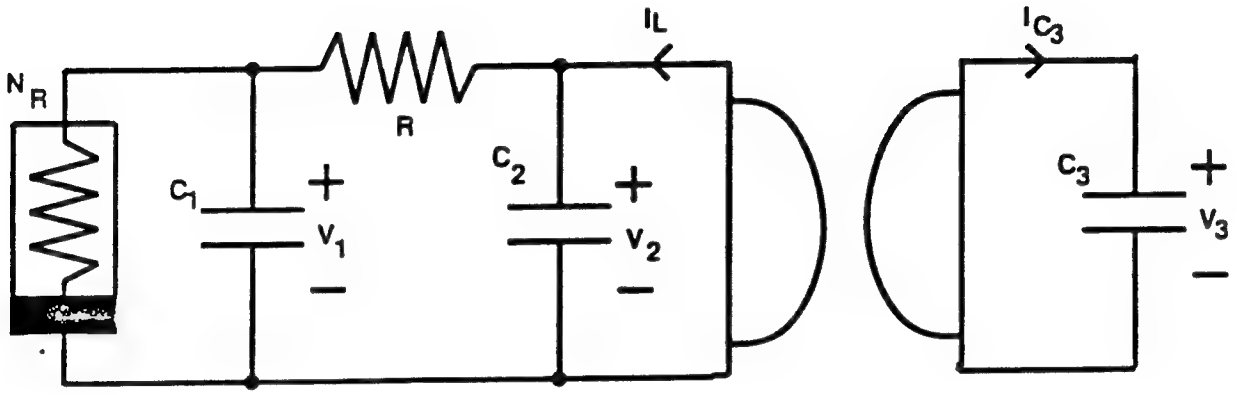


Figure 10

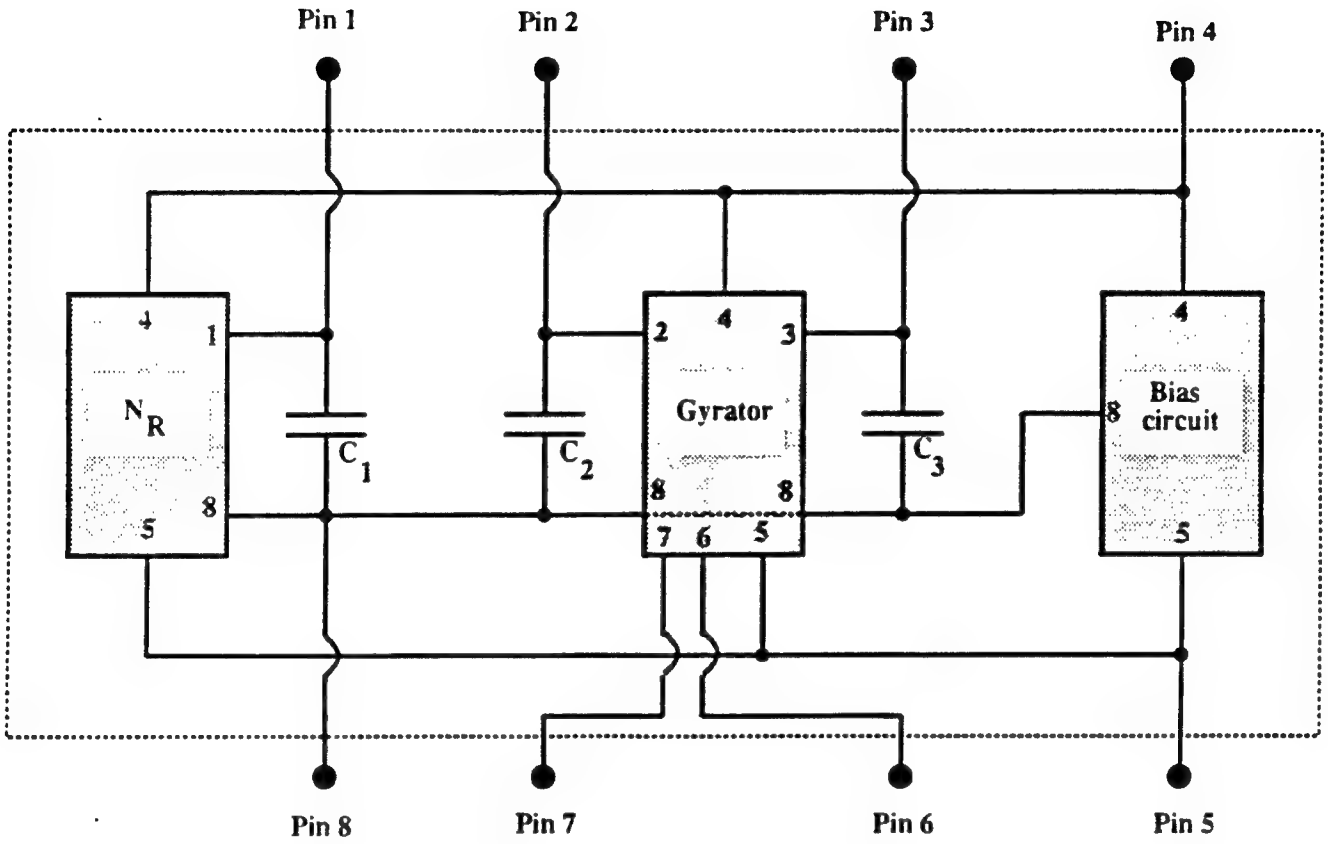

Pin 1
Pin 2
Pin 3
Pin 4
4
1
N_R
5
8
C_1
C_2
2
4
3
Gyrator
8
8
7
6
5
C_3
4
8
Bias
circuit
5
Pin 8
Pin 7
Pin 6
Pin 5

Figure 11

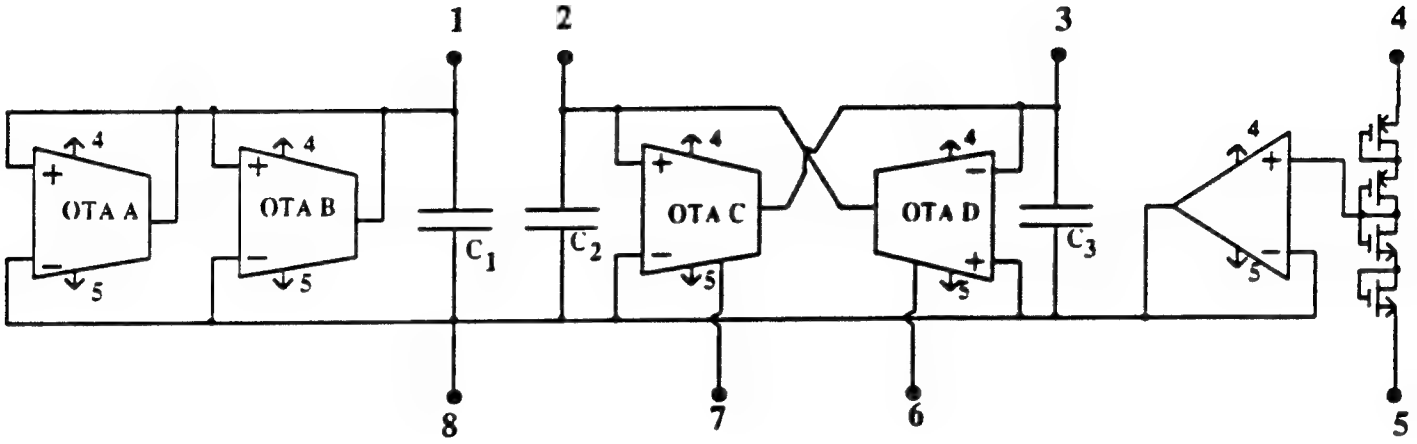


Figure 12

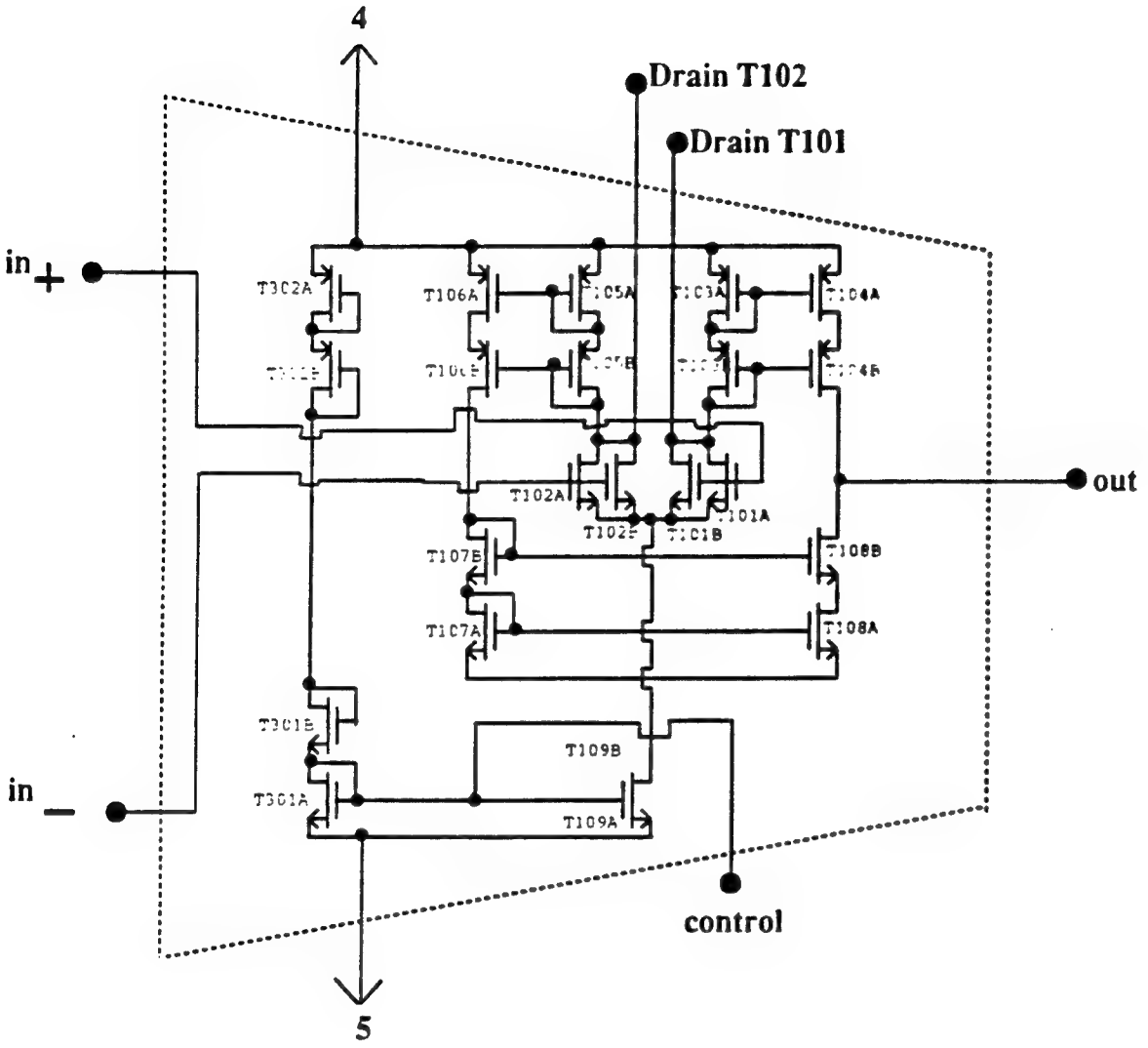

4
Drain T102
Drain T101
in +
in −
out
control
5
T302A
T106A
T105A
T103A
T104A
T102A
T101A
T102B
T101B
T107B
T108B
T107A
T108A
T301B
T109B
T301A
T109A

Figure 13

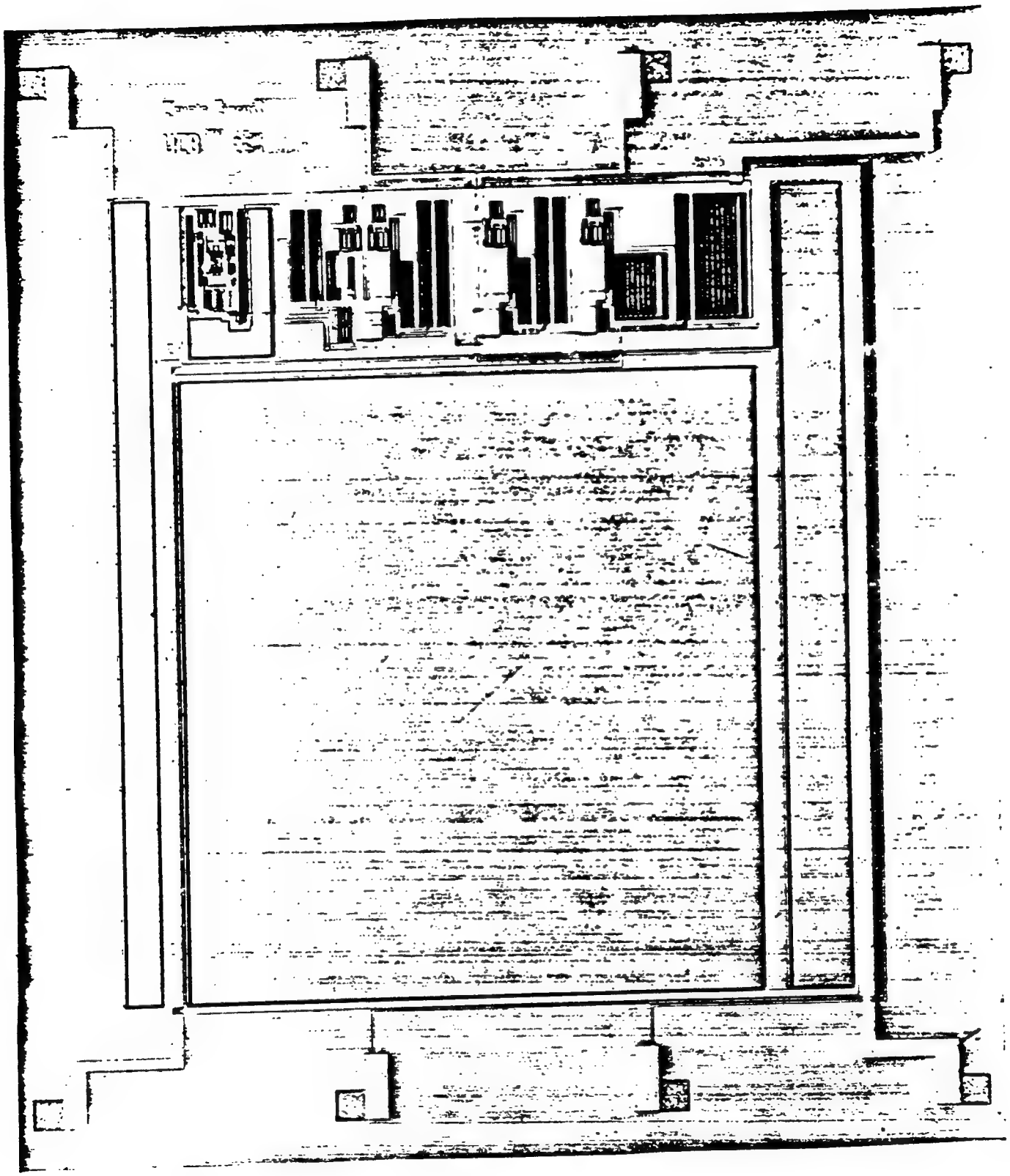


Figure 14

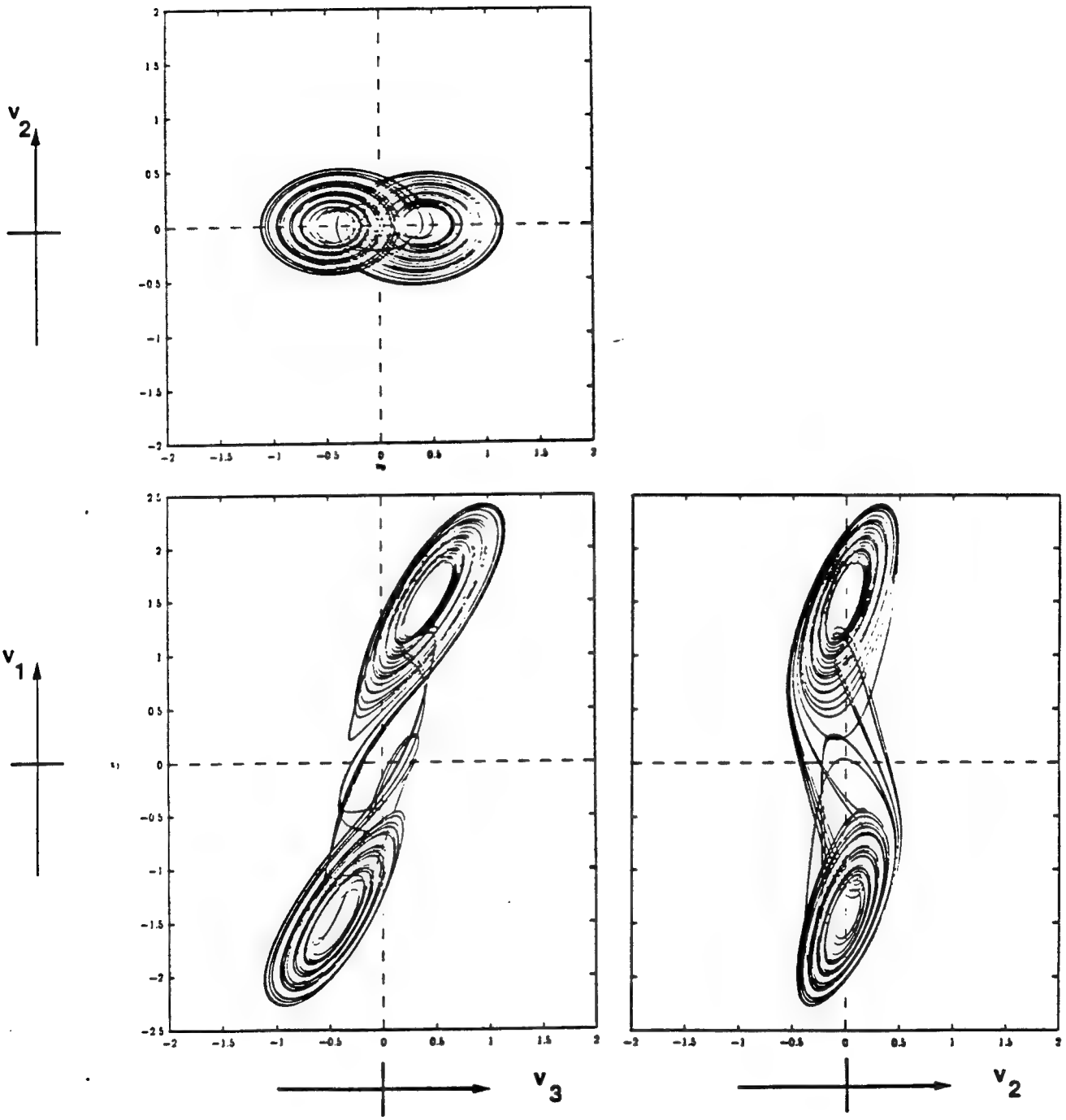


Figure 15

Extended Abstracts of the 1993 International Conference on Solid State Devices and Materials, Makuhari, 1993, pp. 137-139

Invited

Silicon-on-Insulator for High Speed ULSI

Chenming HU

Department of Electrical Engineering and Computer Science
University of California, Berkeley, 94720 USA

ABSTRACT

Bulk CMOS technology scaling can not sustain the historical rate of speed increase. A realistic target for SOI delay and power reductions are 40% and 30%, independent of scaling, mostly through capacitance reduction. Denser isolation allows more compact layout and easy integration of different high speed (E/D NMOS), low power (CMOS) and analog (bipolar, grounded-body CMOS) devices. Silicon device speed record (13 ps at 1.5V, 300K) has been set with SOI E/D NMOS. Leakage current due to steady state and transient floating-body induced threshold lowering (FITL) is a difficult device issue.

The Trend of Bulk Silicon Technology Scaling

The importance of electronics in the economic, ocial and even political development throughout the world will all but guarantee continued rises in circuit integration density and speed. It is less clear if bulk silicon technology can meet the historical trend of speed improvement. A recent study suggests that the speed trend can not be sustained [1]. I_{dsat} per unit channel width ceases to increase with technology scaling beyond the 0.5μm technology. Even when we examine the high-speed scenario, where Vcc reduction is delayed as much as reliability consideration might allow [1], I_{dsat} still ceases to increase. The unpleasant consequence on circuit speed is shown in (Fig. 1). Instead of the historical speed doubling every two generations, designers will need to work with speed doubling every four generations.

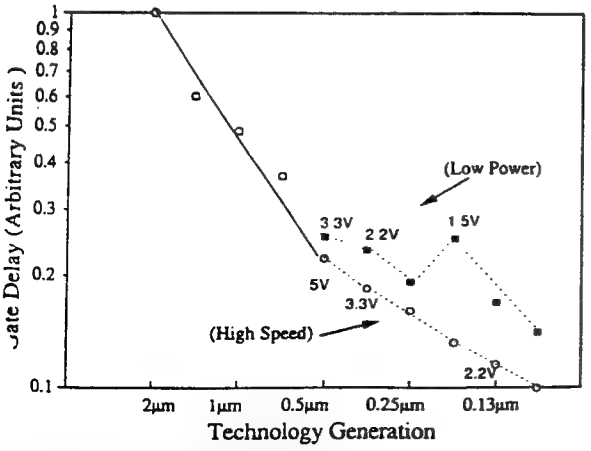


Fig.1. Bulk CMOS scaling can not sustain the rate of speed doubling every 2 generation beyond 0.35μm technology [1].

Capacitance Reduction with SOI

The most often cited advantage of SOI technology is higher speed due to reduction of junction capacitance because of the buried oxide. Comparison of bulk and SOI circuit power consumption provides the most direct data [2]. The ratio of power, $P = f \cdot C \cdot V_{dd}^2$, is equal to the ratio of circuit capacitance. Both data and calculations shown in (Fig. 2) suggest that SOI circuits have approximately two third the capacitance of bulk circuits.

		C (SOI)	C (bulk)	C (SOI) / C (bulk)
Active Gate (F.O.=1)	C_{GT}	36.6 fF	37.6 fF	0.97
N+ Junction (1 drain)	$C_{J(N)}$	9.5 fF	18.9 fF	0.50
P+ Junction (1 drain)	$C_{J(P)}$	7.6 fF	21.6 fF	0.35
Polysilicon (10 μm^2)	C_{POLY}	0.43 fF	0.98 fF	0.44
1st Aluminum (1mm)	C_{1AL}	72.6 fF	123.2 fF	0.59
2nd Aluminum (1mm)	C_{2AL}	63.9 fF	98.4 fF	0.65

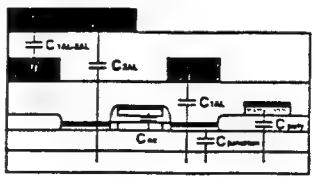


Fig. 2. Comparison of circuit power consumption has confirmed that typical circuit capacitance is reduced to $^2/_3$ of the bulk circuit. Buried oxide is 500nm thick [2].

We expect this capacitance advantage to remain relatively constant independent of scaling. Buried oxide needs to be "electrically" thicker than or physically as thick as the depletion region under the source/drain.

Subthreshold Current and Floating-body Induced Threshold Lowering (FITL)

There are three components of MOSFET leakage current [1]. One is bulk leakage often referred to as puncthrough. SOI eliminates this leakage path easily. The second component is a surface leakage component known as drain-induced barrier lowering, V_T lowering, or short channel effect, $I_d(V_g=0) \propto 10^{-V_t/S}$ SOI offers an opportunity to bring S close to the limit of 2.3kT or 60mV/decade through the use of fully-depleted thin-film SOI [3]. Unfortunately, as V_{ds} increases, drain-body junction leakage and gate-induced drain leakage [4] cause holes (in the case of NMOSFET) to flow into the floating body. This raises the body potential and hence lowers V_T and increases the leakage (Fig. 3) independent of channel length. This can happen even when the silicon film is fully depleted. This floating-body induced leakage is a very serious and difficult problem, especially when one considers transient V_T drift.

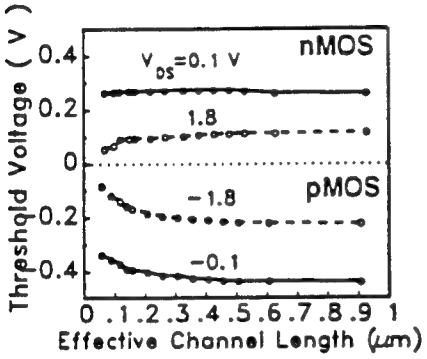


Fig. 3. Floating-body induced threshold lowering (FITL) lowers V_T and raises subthreshold leakage at high V_{ds} even in long-channel devices. Gate oxide is 4.2nm [5].

There are several potential solutions — raise V_T to allow a margin, provide a contact to the body, or make body/source "leaky". We believe there is an important device design concept — use light body doping so that there is minimal potential variation across the silicon film thickness. This "uniform barrier" design will minimize the barrier against hole flow into the source for a given barrier against electron flow into the channel (the subthreshold current).

Enhanced MOSFET Current?

Although reports on SOI devices typically show lower I_{dsat} than bulk devices of the same oxide and channel dimensions, SOI MOSFET can potentially produce larger I_{dsat} than bulk device as shown in (Fig. 4) [6].

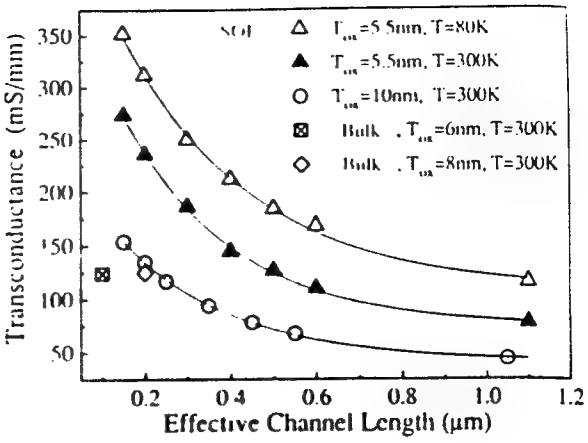


Fig. 4. SOI MOSFET can provide larger current than bulk devices due to reduction in V_T and bulk charge is S/D resistance is not excessive. The 5.5nm gate SOI PMOSFETs produced the highest tranconductance ever reported [6].

The most important reason for SOI's larger I_{dsat} in future low V_{dd} operation is the possibility of lower V_T [1] — if floating-body induced V_T lowering can be controlled. Otherwise, FITL still leads to an effective reduction in V_{Te} of about 0.15V at high V_{ds} and enhanced I_{dsat} in steady state [6]. Finally, reduced bulk charge effect [7] can increase I_{dsat} by around 10% [6] if the buried oxide is effectively much thicker than the bulk depletion region thickness. On the other hand self heating and thin SOI's higher S/D resistance reduce I_{dsat}.

Overall, we expect about the same I_{dsat} in SOI and bulk MOSFET's, with about 10% advantage toward SOI especially at very low V_{dd}, with thicker buried oxide and salicided S/D.

Reliability and Technology Issues

In spite of high dislocation density and metal impurity concentrations, SIMOX as well as bonded SOI materials appear to be capable of producing bulk quality gate oxide [8]. Hot carrier reliability is compromised due to charge trapping in the buried oxide. However effective graded LDD can be produced without trade-off with junction depth. Adequate hot electron reliability is predicted.

More than speed, leakage and reliability issues. Manufacturability will likely decide SOI's future. In this respect, SOI has several advantages in isolations, latch-up, shallow junction, contact formation, layout density, etc. It is worth noting that there are novel and intriguing SOI material and device ideas. One example would produce dense, vertical double-gated thin SOI devices using a bulk silicon substrate as the starting material [9].

Conclusion and Discussion

A critical review suggests that bulk technology scaling can not sustain the historical rate of speed increase. SOI reduces circuit capacitance by 30%, and

can potentially increase MOSFET current by 10% through reduction in V_T and bulk change. Ease of device isolation allows SOI technology to integrate CMOS, complementary BJT, E/D NMOS [10], high voltage device and analog MOSFETs with body contacts. Fastest silicon circuit delay record has been set with SOI E/D NMOS (13 ps at 1.5V and 300K) [11] (Fig. 5). Highest PMOSFET transconductance record has been set with SOI technology (Fig. 4). Manufacturability advantages may favor SOI as the main-stream technology beyond the 0.15 μm technology. Moderately thin (limited by S/D resistance) fully depleted SOI on moderately thick (limited by self heating) buried oxide is the most attractive arrangement. "Uniform barrier" design is proposed to minimize floating-body induced threshold lowering (FITL).

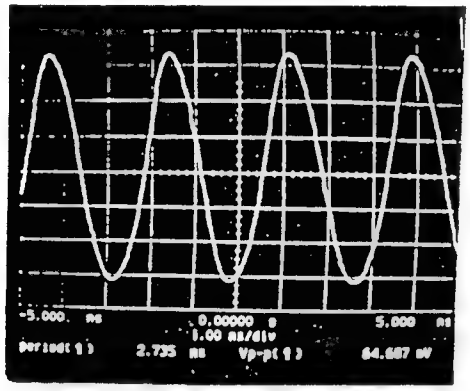


Fig. 5. Fastest silicon transistor speed record, 13ps, has been set with E/D NMOS on SOI technology. 101 inverter ring oscillator, V_{dd}=1.5V, T_{ox}=7nm, T_{si}=50nm, *Leff* = 0.15μm, 300K [11].

ACKNOWLEDMENT

My understanding of the SOI issues has been greatly aided by discussions with Prof. Ping Ko and the research of Dr. Steve Parke, Dr. Jian Chen, Mr. Farib Assaderaghi and Mr. H.J. Wann. Research is partially sponsored by SRC, Texas Instruments, and JSEP under contract F49620-93-C-0014.

REFERENCES

[1] C. Hu "Future CMOS Sealing and Reliability," Proc. of the IEEE, May 1993.

[2] Y. Yamuguchi, et al., "A High-Sped 0.6 μm 16K CMOS Gate Array on a Thin SIMOX Film," IEEE Trans. on Electron Devices, Jan. 1993, pp. 179-186.

[3] J-P. Colinge, "Subthreshold slope of thin film SOI MOSFET's," IEEE Trans. Electron Device Letters, Sept. 1988, pp. 274-276.

[4] J. Chen, et al., "The Enhancement of Gate-Induced-Drain Leakage, Current in SOI MOSFET and Its Application in Measuring Bipolar Current Gain," IEEE Trans. Electron Device Letters, Nov. 1992, pp. 572-574.

[5] G. G. Shahidi, et al., "A Room Temperature 0.1μm CMOS on SOI," Symp. on VLSI Technology Digest, May 1993, pp. 27-28.

[6] F. Assaderaghi, "Study of Current Drive in Deep Sub-Micrometer SOI PMOSFET's," Symp. on VLSI Technology, Systems and Applications, Taipei, May 1993, pp. 232-236.

[7] J. C. Sturm, "Increased Drain Saturation Current in Ultra-thin Silicon-on-Insulator Transistors," IEEE Trans. Electron Device Letters, Sept. 1988, pp. 460-463.

[8] W.M. Huang, et al., "VLSI Quality Gate Oxide on Thin," submitted to 1993 IEDM.

[9] D. Hisamoto, et al., "A Fully Depleted Lean-Channel Transistor DELTA," IEEE Trans. Electron Device Letters, Jan. 1990, pp. 36-38.

[10] S. Parke, et al., "A Versatile SOI BiCMOS Technology with Complimentary Lateral BJT's," IEDM, Dec. 1992, pp. 453-456.

[11] J. Chen, et al., "High Speed SOI Technology," IEDM, Dec. 1992, pp. 35-38.

Ion Beam Synthesis of SiGe Alloy Layers

by

Seongil Im

B.S. (Yonsei University in Seoul, Korea) 1984
M.S. (Yonsei University in Seoul Korea) 1986

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering-Materials Science

and Mineral Engineering

in the

GRADUATE DIVISION

of the

UNIVERSITY of CALIFORNIA at BERKELEY

Committee in charge:

Professor Ronald Gronsky
Professor Jack Washburn
Professor Nathan W. Cheung

1994

Abstract

Ion Beam Synthesis of SiGe Alloy Layers

by

Seongil Im

Doctor of Philosophy in

Engineering-Materials Science and Mineral Engineering

University of California at Berkeley

Professor Ronald Gronsky, Chair

A systematic study of the processing procedures required for minimizing structural defects generated during the ion beam synthesis (IBS) of SiGe alloy layers has been performed. The synthesis of 200 nm thick SiGe alloy layers by implantation of Ge ions with an incident energy of 120 keV into <100> oriented Si wafers yielded various Ge peak concentrations after the following doses, $2x10^{16}cm^{-2}$, $3x10^{16}cm^{-2}$, and $5x10^{16}cm^{-2}$. Following implantation, SPE annealing in a nitrogen ambient at 800°C for 1 hour resulted in only slight redistribution of the implanted Ge. Two kinds of extended defects were observed in alloy layers synthesized at doses over $3x10^{16}cm^{-2}$ at room temperature: end-of-range (EOR) dislocation loops and strain-induced stacking faults. The density of EOR dislocation loops was much lower in those alloys produced by liquid nitrogen temperature (LNT) implantation than by room temperature (RT) implantation. Decreasing the implantation dose to obtain 5 at% peak Ge concentration prevents strain relaxation, while those SPE layers with more than 7 at% Ge peak show high densities of misfit-

induced stacking faults. Sequential implantation of C following high dose ($5x10^{16}/cm^2$) Ge implantation (12 at% Ge peak concentration in the layer) brought about a remarkable decrease in density of misfit-induced defects (stacking faults). When the nominal peak concentration of implanted C was greater than 0.55 at%, stacking fault generation in the epitaxial layer was considerably suppressed. This effect is attributed to strain compensation by C atoms in the SiGe lattice. A SiGe alloy layer with 0.9 at% C peak concentration under a 12 at% Ge peak exhibited the best microstructure. The experimental results, combined with a simple model calculation, indicate that the optimum Ge/C ratio for strain compensation is between 11 and 22. The interface between the amorphous and regrown phases (a/c interface) showed a dramatic morphology change during its migration to the surface. The initial <100> planar interface decomposes into a <111> faceted interface, changing the growth kinetics. These phenomena are associated with strain relaxation by stacking fault formation on (111) planes in the a/c interface.

Ronald Gronsky 4/21/94

INTEGRATING RASTA-PLP INTO SPEECH RECOGNITION

Joachim Koehler[1], *Nelson Morgan*[1], *Hynek Hermansky*[1,2], *H Guenter Hirsch*[*], *Grace Tong*[1]

ICSI and U. of California, Berkeley, California[1]
Oregon Graduate Institute, Portland, Oregon[2]
University of Aachen, Germany[*]

ABSTRACT

In previous work, we and others have shown that bandpass filtering of temporal trajectories of simple functions of the critical band spectrum can lead to more robust speech recognizers in the presence of additive and convolutional error. In this study we report results on several mechanisms for incorporating this analysis technique into training, in a way that is consistent with on-line approaches to speech recognition. In particular, we show improved robustness to these forms of degradation for a system that maps the filtered spectral points using a linear regression computed from results of the different transformations.

1. INTRODUCTION

It has been demonstrated [1][2] that bandpass filtering of temporal trajectories of the critical-band spectrum (when it has been processed by a nonlinear transformation) is efficient in alleviating some harmful effects of both additive and convolutional noise. While the technique appeared to be effective, it raised two new problems:

1. The optimal form of the nonlinearity is dependent on the noise level. Thus, the noise power needs to be estimated for the analysis.
2. Since, depending on the estimated noise level, a different compressive nonlinearity may be applied in the analysis, the result is dependent on the noise level. In a sense this is a trade of a deterministic source of variance (the different nonlinearities used) for a stochastic one (the actual additive or convolutive noise).

Previous work [1] simply estimated the noise power from the non-speech part of the signal to address the first problem. The second problem was addressed by using multiple templates derived from the clean speech using a range of nonlinearities corresponding to the range of expected noise levels.

In the current work we estimate noise without requiring explicit speech detection. Further, we investigate three different techniques for compensating for the effect of the variable nonlinearity. The RASTA models derived for recognition need to match the models derived during training. This was always true for the early forms of RASTA in which the nonlinearity was fixed (a logarithm), but is nontrivial for a nonlinearity whose value is dependent on an adaptively determined parameter (noise level).

2. BACKGROUND

The basic idea of RASTA processing is to filter the temporal trajectories of speech parameters (e.g., critical band values) after they have been transformed by a static nonlinearity that (ideally) converts the major sources of environmental interference into an additive component. Over the last year we have been experimenting with a parameterized family of functions

$$Y_i = \log(1 + JX_i) \qquad (1)$$

where i is the critical band number.

For large values of JX_i, this function is close to logarithmic, while for small values it is close to linear. Experiments reported in [1][2] showed that the optimal value for J is dependent on the instantaneous noise power. To estimate this noise power, we use an approach developed by Hirsch[3] which uses the position of the principal mode of the histogram of energy in each frequency band as the noise power estimate for the band. The subband estimates are currently combined for a robust estimate of the total noise power. This noise estimation technique does not require any speech pause detection.

Though the overall processing has been shown to provide some robustness, a drawback remains: the choice of different J values, as required by differing noise conditions, generates different spectral shapes and dynamics of the spectra. This means that the training system must contend with a new source of variability due to the change in processing strategy that is adaptively determined from the data. The rest of this paper is concerned with the solution of this difficulty.

3. APPROACHES TO HANDLING J VARIABILITY

We have been working on three approaches to handling this variability:

1. Multiple recognizers - several systems can be trained using a different J value for each one. Although clean speech is used for each training, the differing J factors provide a range to include the nonlinear function for cases that will be encountered. In the recognition phase, noise estimation is used to select a J value, and the corresponding recognizer is used. As will be shown, this works well, but several recognizers must be trained.
2. Multiple J values for one recognizer - given enough degrees of freedom in the trained system, training data can be processed for training with a range of plausible values for J. This only requires training a single system, but since this technique effectively increases the size of the training set, it requires more computing and possibly also more parameters in the classifier to account for the added variability.
3. Spectral mapping - the noise-level dependent choice of J introduces a deterministic source of variability into the analysis, which one should be in principle be able to compensate for. To this date, however, we have not determined a satisfactory analytic solution to this problem, and therefore we have decided to apply an empirically derived linear mapping which would transform the spectrum obtained from a J value corresponding to noisy speech to a spectrum processed with a J value for clean speech. In other words, we find a mapping between $log(1 + Jx)$ and $log(1 + J_{ref}x)$. For this approach, we have used a linear regression within each critical band. In principle, this solution reduces the variability due to the choice of J, and so minimizes the effect on the training process.

In the next section we describe experiments to test these three methods.

4. EXPERIMENTS AND RESULTS

We tested our approaches with a standard HMM recognizer which was built with the HMM-Toolkit (HTK)[4]. The recognizer used 10-state word-based HMMs, with 8 emitting states and output probability distributions based on N-Gaussian diagonal covariance matrices. The variances were tied across all HMM states of all models (grand variances). The speech was processed using a 25 ms Hamming window, and then parameterized into 9 PLP-cepstral values. The test database consisted of 13 isolated digits spoken by 200 speakers over dialed-up telephone lines. All words were hand end-pointed. To get enough training data to model the HMMs we divided the set of 200 speakers into 150 speakers for training and 50 speakers for testing. A jackknife procedure was used so that all speakers' data could be tested on, resulting in 4 iterations (no overlap of testing). To balance for the number of parameters, we used 4 mixtures per state for all cases but that of 4 recognizers; for this case we used a single mixture per state (a greater number of mixtures actually didn't substantially change performance for an earlier pilot experiment). To simulate additive noise we synthetically added car noise to the clean (> 20 dB SNR) speech to yield a 10 dB SNR level. Convolutional noise was introduced by filtering the speech with a linear filter simulating the spectral ratio between an electret and carbon microphone. The recognition results are presented in Table 1. The first row gives the results when the environment for train and test phases are identical, and is in some sense a best case scenario for non-RASTA processing; often the testing condition is not available during training. In all other rows the training conditions were always "clean", i.e., the additive and convolutional errors were only applied to test data. The second and third row show the results obtained with PLP and log RASTA processing.

Note that log RASTA (called RASTA here) reduces the error rate for the filtered case but is not effective for additive noise. In this task, RASTA also appears to slightly improve the discriminability between the word classes in the clean case, as in fact one-third of the errors were eliminated with a log RASTA front end (with respect to a PLP front end).

The results using multiple recognizers are shown in the fourth row (J-RASTA-mult). This appears to work reasonably well in comparison with PLP or log RASTA, but there is still a noticeable degradation. In addition, there is a significant performance loss for the clean data.

The next row (J-RASTA-uni) uses one recognizer with data processed using different values of J. This is an HMM version of our multi-template approach [1] and appears to work better than the multiple recognizer technique, both for clean and noisy cases. This case only requires a single recognition step, and so is a fairly straightforward way of incorporating J-RASTA into a recognition system. However, it does still require training with multiple processings of the training data, which increases training time.

The final row shows the results from the linear mapping of filtered critical band values. In this case, J-RASTA-filtered critical band outputs from 10 speakers[1] are used to train linear regression models. We have used 2 coefficients for each of 15 critical bands. Thus, we map the J-RASTA-filtered values for small J (high noise) to the corresponding values for a larger J (low noise). In particular, for each of 3 different values of J (10^{-7}, 10^{-8}, and 10^{-9}), we compute a mapping

$$W_{iJ} = c_1 + c_2 Y_{iJ} \qquad (2)$$

where Y_{iJ} is the J-RASTA-filtered output for critical band i, and the coefficients are determined to minimize the mean-squared error between W_{iJ} and $Y_{iJ_{ref}}$.

The recognizer was trained with clean speech processed with $J = 10^{-6}$, and during recognition the optimal value of J was determined by a local estimate of noise level for the isolated digit. Then the J-RASTA-filtered critical band outputs were linearly mapped using the regression coefficients computed earlier. The performance of this method appears to be quite good. In particular, the score for the clean case is essentially the same as for RASTA (in this case actually better than for PLP), while the mapping approach for the degraded cases are better than for the other approaches (roughly equivalent to the J-RASTA-uni approach). Unlike the other approaches, recognition and training are both the same as for log RASTA, as only a simple deterministic mapping is required in the front end.

[1] Nine of these speakers were independent of the test set; the tenth was one of the 200 speakers in the final testing.

rec. env.	no conv noise		conv noise	
SNR in dB	clean	10	clean	10
PLP same env	95.0	90.0	92.8	89.9
PLP	95.0	63.0	75.1	49.6
RASTA	96.7	50.0	96.4	59.6
J-RASTA-mult	90.9	76.3	87.5	72.1
J-RASTA-uni	92.2	83.2	91.3	81.1
J-RASTA-map	96.3	84.4	94.3	79.2

Table 1 Recognition Performance in %

5. DISCUSSION

The techniques described here permit incorporation of J-RASTA processing in an HMM-based recognizer, at least for a small vocabulary isolated word recognition task. However, the first two of the three increase training time. The third (linear mapping) approach appears preferable from the data shown here (although the train-with-all J-RASTA-uni approach gives slightly higher performance for one condition). However, we do not have enough experience with this method to know whether the mapping is task or data-dependent.

While these techniques do provide significant robustness to additive and convolutional noise, it is clear that, in comparison to the performance on clean speech, there is a significant increase in error which remains. Aside from the smoothing they provide for fast non-speech events, RASTA techniques only handle the constant (or slowly-varying) components of non-linguistic variation.

We close with some caveats about the use of RASTA. In the 2 years since we first reported some RASTA results on recognition, many sites have experimented with related features. Due to the

many different conditions under which these tests were done, results varied from wonderful success to dismal failure with many cases falling in between. Fortunately, this variance does not appear to have a random cause; we have learned a few things about the use of RASTA in recognition of speech. Some of these points are:

- RASTA increases the dependence of the data on its previous context. Therefore, simple context-independent subword-unit recognizers can be degraded by RASTA. We have seen that RASTA has worked well in tasks with whole word models (such as the one reported here), or in phoneme-based recognizers that used triphones or broad temporal input context (the latter being used for our neural-network recognizers).
- Log RASTA does not address the problem of additive noise. J-RASTA in one of the forms described here appears to be able to handle both additive and convolutional noise reasonably well.
- Some RASTA users have had difficulty with initial conditions. One needs to be aware that RASTA incorporates a filter with a significant memory, and thus is different from the well-established short-term spectral analysis of speech in which each analysis frame is entirely independent of its surroundings. To illustrate this point, we originally had difficulty in the experiments reported here when some test files started off with a non-audio artifact which itself was cut off prior to pattern matching, but whose effect spread well into the useful part of the speech data due to the RASTA processing, degrading the performance.

6. ACKNOWLEDGEMENTS

The more recent forms of these experiments have been partially sponsored (at UC Berkeley) under the Joint Services Electronics Program by contract number F49620-93-C-0014. We also acknowledge continuing support from the International Computer Science Institute.

REFERENCES

[1] Hermansky, H., Morgan, N., Hirsch, H.G.: " Recognition of speech in additive and convolutional noise based on RASTA spectral processing", *IEEE Proc. ICASSP'93*, pp. 83-86, 1993

[2] Morgan, N., Hermansky, H.: "RASTA extensions: Robustness to additive and convolutional noise ", *Proc. Workshop on Speech Processing in Adverse Conditions*, Cannes, France, November 1992

[3] Hirsch, H.G.: "Estimation of noise spectrum and its application to SNR-estimation and speech enhancement", *Technical Report TR-93-012*, ICSI, 1993

[4] Woodland, P., and Young, S.: "The HTK Tied-State Continuous Speech Recognizer", Eurospeech '93, pp. 2207-2210, 1993

A Differential Method for Computing Local Shape-From-Texture for Planar and Curved Surfaces

Jitendra Malik Ruth Rosenholtz

Department of Electrical Engineering and Computer Science
University of California at Berkeley, Berkeley, CA 94720
email: malik@cs.berkeley.edu, rruth@robotics.eecs.berkeley.edu

Abstract

We model the texture distortion at a point in any particular direction on the image plane as an affine transformation and derive the relationship between the parameters of the affine transformation and the surface shape and orientation. We use a technique for estimating affine transforms between nearby image patches which is based on solving a system of linear constraints derived from a differential analysis. It is not necessary to explicitly identify texels or make restrictive assumptions about the nature of the image texture like isotropy. We have developed two different algorithms for recovering surface orientation and shape based on the estimated affine transforms in a number of different directions. The first is a simple linear algorithm based on singular value decomposition. The second is based on nonlinear minimization of a least squares error criterion. Experimental results are presented on images of planar and curved surfaces under perspective projection.

1 Introduction

Traditionally, researchers have formalized the notion of texture gradient as that of finding the gradient of certain scalar valued functions such as foreshortening, area, density, compression or scaling. The mathematical relationship between these gradients and scene geometry has been developed both for planar surfaces[16] and curved surfaces [8]. The main difficulty with the use of texture gradients is that it has proven difficult to develop algorithms for estimating the individual texture gradients that do not rely either on explicit texel identification, e.g. Blostein and Ahuha [3], or on restrictive assumptions about the nature of the surface texture such as isotropy. Furthermore, Gårding has shown that the simple distortion gradients do not contain enough information for measurement of complete local surface curvature e.g. sign of Gaussian curvature.

Another major family of approaches to shape from texture in the computer vision literature is based on making some *a priori* assumption about the texture. The assumption most often used is that of isotropy or weak isotropy of the texture [18, 5, 4, 2]. Under projection, the texture will not generally appear isotropic, and thus they use the deviation from isotropy in the projection to infer 3D shape and orientation. There are two major weaknesses of such an approach: (1) It cannot deal with directional texture such as grass, fabrics, etc. (2) It makes only partial use of available information, e.g. it does not exploit the change in size of the projected texture.

The other assumption which has been used in the literature is that of homogeneity, i.e. that the texture pattern has constant area or density[9, 1, 10, 17, 14]. This is a more reasonable assumption for natural textures, and our first criticism doesn't apply. However, this assumption is too weak–it fails to exploit the systematic change in shape of the texture elements.

Obviously, some assumption about the texture is necessary, otherwise what we are seeing could in fact just be a particular pattern of reflectance changes on a flat surface (as in a Renaissance painting). We will assume that the texture is the same at different points on the surface in the scene.

We believe that the natural way to model the texture distortion *locally* is as a 2-D affine transformation between neighboring image patches. We find the affine transformations between two image patches using a differential method (see [12]). In section 2 we derive the relationship between the texture distortion map and all five surface orientation (slant and tilt) and shape (3 parameters: principal curvatures and directions) of the surface, by exploiting previous math-

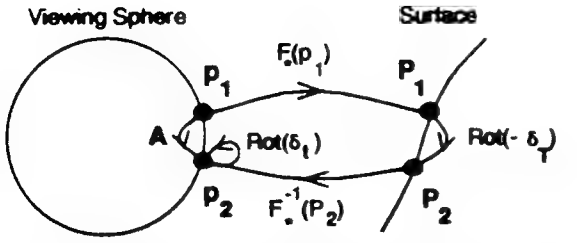


Figure 1: **Determining the affine transformation, A, between the texture at point $\mathbf{p}_1$ and the texture at point $\mathbf{p}_2$.**

ematical analysis of texture gradients on curved surfaces by Gårding [8].

In Section 3 of the paper, we develop two algorithms for recovering surface orientation and shape based on the estimated affine transforms in a number of different directions. The second algorithm will give us estimates of all five shape parameters. We know of no previous researchers who have estimated curvature parameters, though several researchers have derived equations relating the local shape parameters to various texture gradients [8, 10]. We present simulation results on a number of examples of planar and curved surfaces.

2 Relationship between the texture distortion map and 3D shape

We argued previously that instead of studying texture gradients, one should model texture distortion in a particular direction on the image plane as an affine transformation. This section will develop the relationship between the parameters of this affine transformation and the surface shape and pose, for spherical perspective projection.

Figure 1 depicts the situation. We wish to find the matrix, A, which represents the affine transformation between the spherically projected texture at point $\mathbf{p}_1$ and the projected texture at a nearby point $\mathbf{p}_2$. This matrix will be a function of the local orientation and shape parameters. The orientation parameters are σ, the slant of the surface, and $\mathbf{t}$, the direction of tilt of the surface. The shape parameters are $r\kappa_t$, $r\kappa_b$, and $r\tau$, where κ_t is the normal curvature in the tilt direction, κ_b the normal curvature in the perpendicular direction, and τ is the geodesic torsion. The Gaussian curvature is $K = \kappa_t\kappa_b - \tau^2$. The variable r is the distance from the center of the viewing sphere to the given point on the surface. Note that the inseparability of the distance, r, and the curvature parameters is inherent to the problem. The image of a surface S at distance r is indistinguishable from that of a k scaled copy (for which the curvatures will be $1/k$ of S) at a distance of kr.

To find the affine transformation, we first backproject from the point $\mathbf{p_1}$ on the viewsphere to the corresponding point $\mathbf{P_1}$ on the surface, using the map $\mathbf{F_*(p_1)}$. Let $\mathbf{t}$ be a unit vector in the tilt direction for some point $\mathbf{p}$. Then if $\mathbf{p}$ is the unit normal to the viewing sphere in the direction of the surface, let $\mathbf{b} = \mathbf{p} \times \mathbf{t}$. Then $(\mathbf{t}, \mathbf{b})$ forms an orthonormal basis for the tangent plane to the viewing sphere at point $\mathbf{p}$. $\mathbf{t}$ and $\mathbf{b}$ backproject to form the basis $(\mathbf{T}, \mathbf{B})$ on the tangent plane of surface at the backprojection of point p. We will write the backprojection map at point $\mathbf{p_1}$ in terms of the bases $(\mathbf{t_1}, \mathbf{b_1})$ and $(\mathbf{T_1}, \mathbf{B_1})$.

We assume that the texture is constant over the surface, so the transformation between points $\mathbf{P_1}$ and $\mathbf{P_2}$ is only the rotation, by some angle δ_T, between the two bases $(\mathbf{T_1}, \mathbf{B_1})$ and $(\mathbf{T_2}, \mathbf{B_2})$. In [13] we show that the rotation between the bases is

$$\delta_T = (\frac{r\tau}{\sin\sigma})\Delta t + (\sin\sigma + \cos\sigma\cot\sigma + r\kappa_b\cot\sigma)\Delta b \tag{1}$$

as $\Delta t, \Delta b \to 0$. The texture on the surface undergoes rotation by $-\delta_T$.

Next, we project back onto the viewing sphere, using the matrix $F_*^{-1}(\mathbf{P_2})$. This puts us back on the viewing sphere, but in the $(\mathbf{t_2}, \mathbf{b_2})$ basis, not the original $(\mathbf{t_1}, \mathbf{b_1})$ basis. We must convert between these bases by rotating by the angle between the tilt vectors, δ_t. As we show in [13],

$$\delta_t = \frac{r\tau}{\cos\sigma\sin\sigma}\Delta t + (\frac{1}{\sin\sigma})(\cos\sigma + r\kappa_b)\Delta b \tag{2}$$

as $\Delta t, \Delta b \to 0$. Thus we have (see [13])

$$\begin{aligned} A &= \mathrm{Rot}(\delta_t)F_*^{-1}(\mathbf{P_2})\mathrm{Rot}(-\delta_T)F_*(\mathbf{p_1}) \\ &= \mathrm{Rot}(\delta_t)\cdot\begin{bmatrix} k_m\cos\delta_T & k_m\sin\delta_T\cos\sigma \\ -k_M\frac{\sin\delta_T}{\cos\sigma} & k_M\cos\delta_T \end{bmatrix} \end{aligned} \tag{3}$$

where

$$\begin{aligned} k_m &= 1 + \frac{\nabla m}{m}\circ\begin{bmatrix}\Delta t \\ \Delta b\end{bmatrix} \\ k_M &= 1 + \frac{\nabla M}{M}\circ\begin{bmatrix}\Delta t \\ \Delta b\end{bmatrix} \end{aligned} \tag{4}$$

and

$$\frac{\nabla m}{m} = -\tan\sigma\begin{bmatrix} 2 + r\kappa_t/\cos\sigma \\ r\tau \end{bmatrix} \tag{5}$$

$$\frac{\nabla M}{M} = -\tan\sigma\begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{6}$$

The actual affine transformation we find between the two points will not, in general, be in terms of the $(\mathbf{t}, \mathbf{b})$ basis. Instead, our matrix will be related by a change of basis: $\hat{A} = UAU^{-1}$. The change of basis matrix, U, rotates the standard basis to the $(\mathbf{t}, \mathbf{b})$ basis.

The analysis above assumes that we have spherical projection. In reality, cameras use planar projection. We can convert between the two types of projection by applying the Jacobian of the gaze transformation [6, 13].

3 Shape Recovery: Algorithms and Experimental results

To estimate the *texture distortion map* at a point p, we find the spectrograms for that point and for neighboring points in a number of different directions, $\vec{v_i} = (\Delta t_i \quad \Delta b_i)^T$, around p and get estimates, $\tilde{A}_i$, of the affine transforms, $\hat{A}_i$, for each of these directions using the algorithm developed in the previous section. We have developed two algorithms for recovering surface orientation (slant and tilt) and shape (principal curvatures and directions). We will present experimental results on a number of images of planar and curved objects.

The first shape recovery method is based on singular value decomposition (SVD) of the $\tilde{A}_i$ matrices to estimate k^i_m and k^i_M. By solving two associated systems of linear equations, we obtain estimates of the slant, σ, the tilt direction specified by θ_t, and the shape parameters $r\kappa_t$ and $r\tau$ [12, 13]. Note that for the systems of linear equations to be solvable, it is necessary and sufficient to find affine transforms in two independent directions.

Table 1 shows the results of the algorithm on seven images. For each of these images, we found the affine transform in eight different directions around a given point of the image. Each of these images was created by mapping Brodatz textures on various surfaces. The image marked "noise" is the "wire" image, with added noise of standard deviation 30. The first four surfaces are planar, followed by a cylinder and two spheres. The torsion, τ, is zero for all of the examples.

Complete information about local surface shape requires knowledge of three parameters, and here we have only found two: $r\kappa_t$ and $r\tau$; the third parameter $r\kappa_b$ is left undetermined. This was to be expected as this algorithm is essentially based on factoring the affine transform matrices to obtain the major and minor axis gradients–we know from Gårding that these underspecify shape.

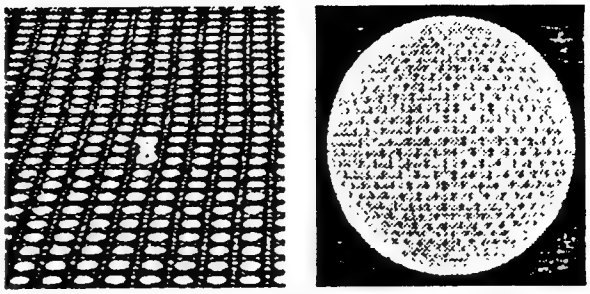


Figure 2: Two example textured surfaces.

	True				*Estimated*			
#	σ	tilt	$r\kappa_t$	$r\kappa_b$	σ	tilt	$r\kappa_t$	$r\tau$
1	69	-90	0	0	58	-91	.95	-.04
2	64	-25	0	0	56	-36	.79	1.2
3	64	-25	0	0	57	-33	.76	.95
4	64	-90	0	0	49	-106	-.14	1.7
5	28	180	6.5	0	42	171	.49	.03
6	39	180	6.6	6.6	80	166	-.26	.46
7	40	180	6.6	6.6	79	174	-.28	.43

Table 1: True and Estimated Surface Parameters: Method 1

For the planar cases, we get consistent underestimation of slant, but in general the results are good. The algorithm does quite well even in the presence of noise. We get good results on the "straw" planar surface, even though this texture does not satisfy the isotropy assumption commonly used by other researchers. The tilt and some of the slant estimates for the curved surfaces were also reasonable, but the curvature estimates were not accurate enough to be usable. If this algorithm were to be used for shape estimation by itself, the recommended strategy would be to obtain slant and tilt estimates at a large number of points and then fit a smooth surface consistent with these estimates. Differentiating this fitted surface yields the desired shape parameters.

A better approach is to use our second algorithm which gives much more accurate orientation and shape estimates *locally* without any need for a surface fitting stage. The second algorithm is based on finding the orientation and shape parameters which minimize the sum of squared errors between the predicted and empirically measured entries of the affine transformation matrices, i.e., we wish to minimize the following error

Image	Estimated					
	σ	tilt	$r\kappa_t$	$r\kappa_b$	$r\tau$	χ^2_{min}
1	71	-92	-.06	.00	-.03	.031
2	63	-20	.22	.00	-.03	.027
3	61	-18	.32	-.01	-.10	.027
4	60	-73	-.13	.29	-.27	.010
5	26	180	3.2	-.13	-.15	.003
6	40	179	6.0	5.8	.67	.002
7	35	182	8.0	5.6	.95	.002

Table 2: True and Estimated Surface Parameters. Method 2

Image	Error Bounds				
	σ	tilt	$r\kappa_t$	$r\kappa_b$	$r\tau$
1	$\pm$1.3	$\pm$2.6	$\pm$.07	$\pm$.26	$\pm$.09
2	$\pm$1.5	$\pm$2.4	$\pm$.08	$\pm$.19	$\pm$.08
3	$\pm$1.5	$\pm$2.4	$\pm$.08	$\pm$.19	$\pm$.08
4	$\pm$3.3	$\pm$4.3	$\pm$.22	$\pm$.35	$\pm$.16
5	$\pm$13.0	$\pm$5.2	$\pm$5.4	$\pm$.79	$\pm$.76
6	$\pm$11.6	$\pm$3.8	$\pm$4.5	$\pm$1.9	$\pm$.63
7	$\pm$11.4	$\pm$3.8	$\pm$4.4	$\pm$1.8	$\pm$.62

Table 3: Error Bounds on True Parameters

function:

$$\chi^2(\sigma, \theta_t, r\kappa_t, r\kappa_b, r\tau) = \sum_{i=1}^{n} \sum_{k=1}^{2} \sum_{l=1}^{2} (\hat{A}_i(k,l) - \bar{A}_i(k,l))^2$$

where $\hat{A}_i(k,l)$ the (k,l)th element of the theoretically predicted matrix $\hat{A}_i$ and is a function of the shape parameters, and $\bar{A}_i(k,l)$ is the (k,l)th element of the empirically measured affine transform matrix $\bar{A}_i$. Ideally, each term in this error sum should be weighted by the inverse of the standard deviation of the measurement error of that particular entry. A specific characterization of the probability distribution of the measurement errors in the entries of the affine transform matrices A_i is not yet available. We expect it to depend on the particular algorithm used for the estimation of the affine transforms. In the absence of a particularly appropriate model, we will proceed on the (convenient!) assumption that these errors are independent and normally distributed with standard deviation Δa.

For minimizing the error function, we just used the gradient descent routine in the *Mathematica* package– there are any of a number of variants such as conjugate gradient, Levenberg-Marquardt that could have been used equivalently. The starting point is provided by the orientation and shape estimate returned by the first algorithm. Table 2 shows the results for the same images as in Table 1. We get improved slant and tilt estimates, and also significantly better estimates of the curvature parameters. We will now determine confidence intervals on the orientation and shape estimates. For more details on the methodology that we follow, the reader is referred to [15], Chapters 14.4 and 14.5.

Let us abbreviate the five geometrical parameters $(\sigma, \theta_t, r\kappa_t, r\kappa_b, r\tau)$ as $g_i, i = 1, 2, \ldots 5$. To obtain confidence intervals on the parameters, one computes the so-called *curvature matrix* $[\alpha]$ which is defined as half of the Hessian of the χ^2 function

$$\alpha_{kl} = \frac{1}{2\Delta a^2} \frac{\partial^2 \chi^2}{\partial g_k \partial g_l}$$

The inverse of this 5×5 curvature matrix is the covariance matrix, C, of the fit, on the assumption of independent, identically normally distributed errors in the entries of the affine transformation matrices. In that case the confidence intervals for the parameters are given by $\sqrt{C_{ii}}$. The confidence interval for parameter g_i, $\pm\delta g_i$ is $\pm\sqrt{C_{ii}}$ for 68 percent confidence, $\pm 2\sqrt{C_{ii}}$ for 95 percent confidence. In Table 3 we give the 68 percent confidence intervals for slant, assuming[1] a measurement error Δa of standard deviation 0.0323. Note that these are intervals surrounding the true parameters. Using this value for the standard deviation, 66 percent of the all of the estimated parameters fall within the 68 percent confidence intervals of the true parameters.

In conclusion, we have presented a method for finding the shape of surfaces locally from texture distortion, modeled as a set of affine transforms in different directions in the image. The advantage of this representation is that it captures *all* the information available locally and does so without any restrictive assumptions. We develop a differential technique for estimating the affine transforms, which can be applied to a number of other vision problems.

Our results demonstrate that local shape-from-texture without any *a priori* assumptions on the texture is a viable module for early vision.

This research was supported by NSF PYI grant IRI-8957274, Xerox, the PATH project, and JSEP contract F49620-93-C-0014. We thank Pietro Perona and

[1] We computed this value as the standard deviation of the errors in the affine transformation matrices for the examples used in this paper. The errors are known because we know the ground truth in these synthetic examples.

Phil Lapsley for useful discussions, and John Oliensis for catching an error in an earlier version.

References

[1] J. Aloimonos, "Shape from texture," *Biological Cybernetics*, Vol. 58, pp. 345-360, 1988.

[2] A. Blake and C. Marinos, "Shape from texture: estimation, isotropy and moments." *Artificial Intelligence*, 45(1990):323-380.

[3] D.Blostein and N. Ahuja, "Shape from texture: integrating texture-element extraction and surface estimation," *IEEE Trans. on PAMI*, 11(12):1233-1251, December 1989.

[4] L.G. Brown and H. Shvaytser,"Surface orientation from projective foreshortening of isotropic texture autocorrelation," *IEEE Trans. on PAMI*, 12(6), June 1990, pp. 584-588.

[5] L.S. Davis, L. Janos, and S.M. Dunn, "Efficient recovery of shape from texture," *IEEE Trans. on PAMI*, 5(5), 1983.

[6] J. Gårding, "Shape from surface markings," Ph.D. Dissertation, Dept. of Numerical Analysis and Computing Science, Royal Institute of Technology, 1991.

[7] J. Gårding, "Shape from texture for smooth curved surfaces in perspective projection," CVAP Technical Report, TRITA-NA-P9203, Royal Institute of Technology, 1992.

[8] J. Gårding, "Shape from texture for smooth curved surfaces," Proc. of Second ECCV, Santa Margherita Ligure, Italy, 1992, pp. 630–638.

[9] K. Ikeuchi, "Shape From Regular Patterns," *J. of Artificial Intelligence*, 22:49-75, 1984.

[10] K. Kanatani and T.C. Chou, "Shape from texture: General Principle," J. of Artificial Intelligence, Vol. 38, pp. 1-48, 1989.

[11] T. Lindeberg and J. Gårding, "Shape from Texture from a Multi-Scale Perspective," Proc. ICCV, Berlin, Germany, 1993, pp. 683-691.

[12] J. Malik and R. Rosenholtz, "A Differential Method for Computing Local Shape-From-Texture for Planar and Curved Surfaces," Proc. CVPR, New York, 1993.

[13] J. Malik and R. Rosenholtz, "A Differential Method for Computing Local Shape-From-Texture for Planar and Curved Surfaces," CS Division Technical Report, UCB-CSD-93-775, UC Berkeley, 1993.

[14] C. Marinos and A. Blake, "Shape from texture: the homogeneity hypothesis," Proc. ICCV, Osaka, Japan, 1990, pp.350-353.

[15] W.H. Press, B.P. Flannery, S.A. Teukolski, W.T. Vetterling, "Numerical Recipes in C," Cambridge University Press, 1988.

[16] K.A. Stevens, "The information content of texture gradients," *Biological Cybernetics*, Vol. 42, pp. 95-105, 1981.

[17] B. Super and A. Bovik, "Shape-from-Texture by Wavelet-Based Measurement of Local Spectral Moments," Proc. CVPR, Champaign-Urbana, Illinois, 1992, 296-301.

[18] A.P. Witkin, "Recovering surface shape and orientation from texture," J. of Artificial Intelligence, Vol. 17, pp. 17-45, 1981.

Recovering Surface Curvature and Orientation From Texture Distortion: A Least Squares Algorithm and Sensitivity Analysis

Jitendra Malik and Ruth Rosenholtz

Dept. of Electrical Engineering and Computer Science,
University of California at Berkeley, Berkeley, CA 94720, USA
malik@cs.berkeley.edu, rruth@robotics.eecs.berkeley.edu

Abstract. Shape from texture is best analyzed in two stages, analogous to stereopsis and structure from motion: (a) Computing the 'texture distortion', and (b) Interpreting the 'texture distortion' to infer the orientation and shape of the surface. We model the texture distortion for a given point and direction on the image plane as an affine transformation and derive the relationship between the parameters of this transformation and the shape parameters. We use non-linear minimization of a least squares error criterion to estimate the shape parameters from the affine transformations, using a simple linear algorithm to obtain an initial guess. Under the assumption that the measurement errors in the affine parameters are independent and normally distributed, we can find error bounds on the shape parameter estimates. We present results on images of planar and curved surfaces under perspective projection. We find all five local shape and orientation parameters with no a priori assumptions about the shape of the surface.

1 Introduction

In its geometric essence, shape from texture is a cue to 3D shape very similar to binocular stereopsis and structure from motion. All of these cues are based on the information available in multiple perspective views of the same surface in the scene. In binocular stereopsis, the two eyes get slightly different views of the same surface; in structure from motion, the relative motion of the observer and the surface generates the different views. To put shape from texture in this framework, consider two nearby patches on a surface in the scene with same (or sufficiently similar) texture. The appearances of the two patches in a single monocular image will be slightly different because of the slightly different geometrical relationships that they have with respect to the observer's eye or camera. We thus get multiple views in a *single* image.

This naturally suggests a two stage framework (1) Computing the 'texture distortion' from the image, and (2) Interpreting the 'texture distortion' to infer the orientation and shape of the scene surface in 3D. The 'texture distortion' is the counterpart in texture analysis of binocular disparity in stereopsis or optical flow in structure from motion. We believe that the natural way to model the

texture distortion *locally* is as a 2-D affine transformation between neighboring image patches. This affine transformation will depend on the direction and magnitude of the vector displacement between the two patches in the image. We find the affine transformations between two image patches using a differential method (see [12, 13]). We will call the map which associates to each direction in the image plane an affine transformation, the ***texture distortion map***. For each point on a smoothly curved textured surface this map is well defined and can be related to surface shape and orientation with respect to the viewer.

In Sect. 3 we derive the relationship between the texture distortion map and the surface orientation (slant and tilt) and shape (principal curvatures and directions). This derivation makes use of previous results due to Gårding [8] and builds on our previous derivation in [12]

In Sect. 4 of the paper, we develop a new algorithm for recovering surface orientation and shape based on the estimated affine transforms in a number of different directions. The method uses nonlinear minimization of a least squares error criterion to estimate the shape parameters. We use a simple linear algorithm based on singular value decomposition of the linear parts of the affine transforms to find the initial conditions for the minimization procedure [13]. This linear algorithm is a slight modification of the work we described in [12]. and we will not describe it in this paper.

Our shape estimation algorithm is arguably optimal in a maximum likelihood sense if the measurement errors in the affine parameters can be assumed to be independent and normally distributed. By studying the Hessian of the error function at the minimum point, one can characterize the confidence intervals of the shape estimates. We present simulation results on a number of examples of planar and curved surfaces. Finally, we will discuss predictions for human perception of shape from texture.

2 Relationship to Previous Work

We review previous shape from texture research in detail in [12, 13]. Much of the previous research in shape from texture has assumed either a homogeneous (i.e. constant area or density)[9, 1, 10, 17, 14], or an isotropic texture[18, 6, 5, 4] in the scene. The isotropy assumption will be incorrect for many textures such as grass, fabric, and bricks. Both of these assumptions allow one to make only partial use of the available information; under the homogeneity assumption one cannot make use of the change in shape of the texture elements, and under the isotropy assumption one cannot make use of the change in size of the elements. Obviously, some assumption about the texture is necessary, otherwise what we are seeing could in fact just be a particular pattern of reflectance changes on a flat surface (as in a Renaissance painting). We will assume that the texture is the same at different points on the surface in the scene. While this implies periodicity for a deterministic pattern, for a texture which is best thought of as a realization of a stochastic process we can formalize this as stationarity under translations. The term ***homogeneity*** is used in the probability and statistical

literature as equivalent to stationarity under translations. A stochastic process is specified by giving the joint distributions of any finite subsets of the variables. The thing to note here is that we can assume that not only the first but also the second (and higher) order statistics are translation-invariant. This is more powerful than assuming, as in previous use of homogeneity in the computer vision literature, that just the first order statistics (e.g. fraction of surface area occupied by texels) are translation invariant. We will be able to exploit changes in shapes of texture elements as well.

In the Sect. 3 we will relate the parameters of an affine transformation between two image patches to five of the local shape and orientation parameters: slant, tilt, and three curvature parameters. Section 4 presents the algorithm for estimating all five shape parameters, with results on synthetic and real images. To our knowledge, this is the first time that direct estimation of curvature parameters from textured images has been demonstrated. While Gårding's [8] analysis dealt with general curved surfaces, his algorithms for estimating shape from distortion gradients permit only the computation of slant and tilt.

3 Relationship Between the Texture Distortion Map and 3D Shape

In this section we will develop the relationship between the parameters of the affine transformation between a pair of images patches and the surface shape and pose. We use perspective projection to a spherical image surface instead of to a planar surface. While there is a 1-1 mapping which relates the two kinds of perspective projection, known as the *gaze transformation*, the relations which follow turn out to be simpler in the spherical case [9, 8].

This section has two parts. The first part is a review of the formalism developed by Gårding[8]–essentially he defines an orthonormal frame field on the image sphere with one of the vectors in the tilt direction. The backprojection map takes on a particularly simple form, and he is able to obtain expressions for the different texture gradients for the general situation of smooth curved surfaces under perspective projection.

In the second part, we exploit this frame field to derive an expression for the affine transformation on the image sphere which relates two neighboring image *patches*.

3.1 The slant-tilt frame field

This section is based on Gårding [8] to which the reader is referred to for proofs of the various assertions. Relevant differential geometry concepts may be found in O'Neill [15].

The basic geometry is illustrated in Fig. 1. A smooth surface S is mapped by central projection to a unit sphere Σ centered at the focal point. The backprojection map F from Σ to S is defined as $F(\mathbf{p}) = \mathbf{r}(\mathbf{p}) = r(\mathbf{p})\mathbf{p}$ where $\mathbf{p}$ is a unit vector from the focal point to a point on the image sphere, and $r(\mathbf{p})$ is the

distance along the visual ray from the focal point through $\mathbf{p}$ to the corresponding point $\mathbf{r} = F(\mathbf{p})$ on the surface S. We consider this map for regions of the surface where the map is not singular by excluding neighborhoods containing the occluding contour. The derivative of the backprojection map F_* maps tangent vectors of Σ at $\mathbf{p}$ to tangent vectors of S at $F(\mathbf{p})$.

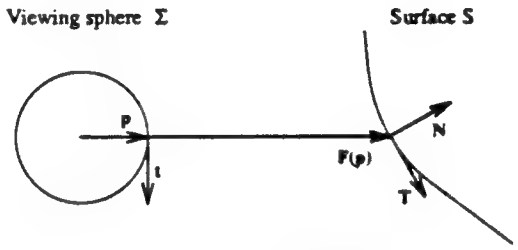


Fig. 1. Local surface geometry.

Define the tilt direction $\mathbf{t}$ in $T_{\mathbf{p}}(\Sigma)$, the tangent plane of the viewing sphere at $\mathbf{p}$, to be a unit vector in the direction of the gradient of the distance function $r(\mathbf{p})$, and the auxiliary vector $\mathbf{b} = \mathbf{p} \times \mathbf{t}$. Then $(\mathbf{t}, \mathbf{b})$ form an orthonormal basis for the tangent plane to the image sphere Σ and together with $\mathbf{p}$ constitute an orthonormal frame field on Σ. Garding shows that $\mathbf{t}$ and $\mathbf{b}$ backproject to orthogonal vectors $F_*(\mathbf{t}) = r'\mathbf{p} + r\mathbf{t}$ and $F_*(\mathbf{b}) = r\mathbf{b}$ in the tangent space $T_{F(p)}(S)$. Dividing these vectors by their lengths gives us an orthonormal basis $(\mathbf{T}, \mathbf{B})$ of the tangent space of the surface at $F(\mathbf{p})$. The vectors $\mathbf{T}, \mathbf{B}$ along with the unit normal to the surface $\mathbf{N} = \mathbf{T} \times \mathbf{B}$ constitute an orthonormal frame field on the surface. The slant angle σ is defined to be the angle between the surface normal $\mathbf{N}$ and the viewing direction $\mathbf{p}$, so that $\cos\sigma = \mathbf{N}.\mathbf{p}$.

The shape of the surface is captured in the *shape operator*, which measures how the surface normal $\mathbf{N}$ changes as one moves in various directions in the tangent space of the surface $T_{F(p)}(S)$. One can represent the shape operator in the $(\mathbf{T}, \mathbf{B})$ basis as

$$\begin{bmatrix} -\nabla_T \mathbf{N} \\ -\nabla_B \mathbf{N} \end{bmatrix} (\mathbf{p}) = \begin{bmatrix} \kappa_t & \tau \\ \tau & \kappa_b \end{bmatrix} \begin{bmatrix} \mathbf{T} \\ \mathbf{B} \end{bmatrix} \tag{1}$$

where κ_t is the normal curvature in the $\mathbf{T}$ direction, κ_b the normal curvature in the $\mathbf{B}$ direction and τ is the geodesic torsion. The determinant of the operator gives the Gaussian curvature $K = \kappa_t\kappa_b - \tau^2$, and half the trace is the mean curvature $H = (\kappa_t + \kappa_b)/2$.

Gårding goes on to obtain expressions for the derivatives of the $(\mathbf{p}, \mathbf{t}, \mathbf{b})$ frame field on Σ expressed in terms of the frame field itself. One can also define the derivatives of the frame field $(\mathbf{N}, \mathbf{T}, \mathbf{B})$ on S with respect to $(\mathbf{p}, \mathbf{t}, \mathbf{b})$ by first pulling back these fields from the surface S to Σ.

Using the linearity of the derivative, we can compute $\nabla_{\mathbf{v}}\mathbf{t}$ for an arbitrary

vector $\mathbf{v} = \Delta t\,\mathbf{t} + \Delta b\,\mathbf{b}$ in the tangent space.

$$\nabla_{\mathbf{v}}\mathbf{t} = -\Delta t\,\mathbf{p} + \frac{r\tau}{\cos\sigma\sin\sigma}\Delta t\,\mathbf{b} + (\frac{1}{\sin\sigma})(\cos\sigma + r\kappa_b)\Delta b\,\mathbf{b} \qquad (2)$$

where r is the distance to the object from the center of the viewing sphere.

In the next section, we will also need the derivative of the $\mathbf{T}$ vector field. We can compute $\nabla_{\mathbf{v}}\mathbf{T}$ for an arbitrary vector $\mathbf{v} = \Delta t\,\mathbf{t} + \Delta b\,\mathbf{b}$ in the tangent space[13]:

$$\nabla_{\mathbf{v}}\mathbf{T} = (\frac{rk_t}{\cos\sigma}\Delta t + r\tau\Delta b)\mathbf{N} + (\frac{r\tau}{\sin\sigma})\Delta t\,\mathbf{B} + \frac{1}{\sin\sigma}(1 + r\kappa_b\cos\sigma)\Delta b\,\mathbf{B} \qquad (3)$$

3.2 Affine Transformations on the Image Sphere

Figure 2 depicts the situation. We wish to find the matrix, A, which represents

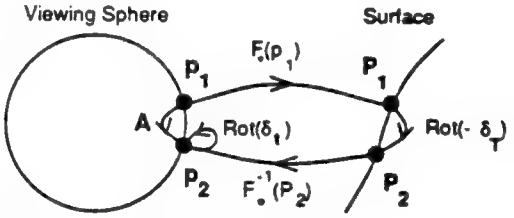


Fig. 2. Determining the affine transformation, A, between the texture at point p_1 and the texture at point p_2.

the affine transformation between the spherically projected texture at point $\mathbf{p_1}$ and the projected texture at a nearby point $\mathbf{p_2}$. This matrix will be a function of the local orientation and shape parameters. The orientation parameters are σ, the slant of the surface, and $\mathbf{t}$, the direction of tilt of the surface. The shape parameters are $r\kappa_t$, $r\kappa_b$, and $r\tau$. The variable r is the distance from the center of the viewing sphere to the given point on the surface. Note that the inseparability of the distance, r, and the curvature parameters is inherent to the problem. The image of a surface S at distance r is indistinguishable from that of a k scaled copy (for which the curvatures will be $1/k$ of S) at a distance of kr.

Our analysis is a differential analysis. We will freely assume that $\mathbf{p_2} - \mathbf{p_1}$ can be modelled as a vector $\mathbf{v} = \Delta t\,\mathbf{t} + \Delta b\,\mathbf{b}$ in the tangent space at the point $\mathbf{p_1}$ and the expressions derived will be true in the limit as $\Delta t, \Delta b \to 0$.

To find the affine transformation, we first backproject from the point $\mathbf{p_1}$ on the viewsphere to the corresponding point $\mathbf{P_1}$ on the surface, using the map $\mathbf{F_*}(\mathbf{p_1})$. Using the basis $(\mathbf{t_1}, \mathbf{b_1})$ on the tangent plane of the image sphere Σ, and $(\mathbf{T_1}, \mathbf{B_1})$ on the tangent plane on the surface S, this map can be represented as

$$F_*(\mathbf{p_1}) = \begin{bmatrix} r/\cos\sigma & 0 \\ 0 & r \end{bmatrix} = \begin{bmatrix} \frac{1}{m_1} & 0 \\ 0 & \frac{1}{M_1} \end{bmatrix}$$

We see that m_1 is the scaling of the texture pattern in the "minor axis," i.e. in the tilt direction, due to projection, and M_1 is the scaling in the "major axis."

We assume that the texture is constant over the surface, so the transformation between points $\mathbf{P_1}$ and $\mathbf{P_2}$ is only the rotation, by some angle δ_T, between the two bases $(\mathbf{T}_1, \mathbf{B}_1)$ and $(\mathbf{T}_2, \mathbf{B}_2)$. To find this angle, we begin by noting that $\mathbf{T_2} = \mathbf{T_1} + \nabla_{\mathbf{v}}\mathbf{T}$ to first order, and hence from equation 3

$$\mathbf{T_2} = \mathbf{T_1} + (\frac{r k_t}{\cos\sigma}\Delta t + r\tau\Delta b)\mathbf{N_1} + (\frac{r\tau}{\sin\sigma})\Delta t\,\mathbf{B_1} + \frac{1}{\sin\sigma}(1 + r\kappa_b\cos\sigma)\Delta b\,\mathbf{B_1} \quad (4)$$

In the right hand side of this equation, the term in the direction of the surface normal $\mathbf{N_1}$ represents a change of the plane of the frame as a whole, and the terms in the direction of $\mathbf{B_1}$ represent a rotation about the surface normal on S. Since $\mathbf{T}, \mathbf{B}$ are unit vectors, we see that

$$\delta_T = (\frac{r\tau}{\sin\sigma})\Delta t + \frac{1}{\sin\sigma}(1 + r\kappa_b\cos\sigma)\Delta b \quad (5)$$

as $\Delta t, \Delta b \to 0$. Note that if the $\mathbf{T}, \mathbf{B}$ basis vectors undergo counterclockwise rotation by δ_T, the texture on the surface undergoes rotation by $-\delta_T$.

Next, we project back onto the viewing sphere, using the matrix $F_*^{-1}(\mathbf{P_2})$. This puts us back on the viewing sphere, but in the $(\mathbf{t_2}, \mathbf{b_2})$ basis, not the original $(\mathbf{t_1}, \mathbf{b_1})$ basis. We must convert between these bases by rotating by the angle between the tilt vectors, δ_t. As in the case of the $\mathbf{T}$ vector field, to find this angle we begin by noting that $\mathbf{t_2} = \mathbf{t_1} + \nabla_{\mathbf{v}}\mathbf{t}$ to first order. Hence from equation 2 we get

$$\mathbf{t_2} = \mathbf{t_1} - \Delta t\,\mathbf{p_1} + \frac{r\tau}{\cos\sigma\sin\sigma}\Delta t\,\mathbf{b_1} + (\frac{1}{\sin\sigma})(\cos\sigma + r\kappa_b)\Delta b\,\mathbf{b_1} \quad (6)$$

As before, the term in the direction of $\mathbf{p_1}$ represents a change of the plane of the frame as a whole, and the term in the direction of $\mathbf{b_1}$ represents a rotation of the frame about the normal. We obtain

$$\delta_t = \frac{r\tau}{\cos\sigma\sin\sigma}\Delta t + (\frac{1}{\sin\sigma})(\cos\sigma + r\kappa_b)\Delta b \quad (7)$$

as $\Delta t, \Delta b \to 0$. Thus we have

$$A = \mathrm{Rot}(\delta_t)F_*^{-1}(\mathbf{p_2})\mathrm{Rot}(-\delta_T)F_*(\mathbf{p_1}) = \mathrm{Rot}(\delta_t)\cdot\begin{bmatrix}\cos\delta_T\frac{m_2}{m_1} & \sin\delta_T\frac{m_2}{M_1}\\ -\sin\delta_T\frac{M_2}{m_1} & \cos\delta_T\frac{M_2}{M_1}\end{bmatrix} \quad (8)$$

Gårding showed that the normalized gradients of the minor and major axis scale factors, in the $(\mathbf{t}, \mathbf{b})$ basis, are

$$\frac{\nabla m}{m} = -\tan\sigma\begin{bmatrix}2 + r\kappa_t/\cos\sigma\\ r\tau\end{bmatrix}, \qquad \frac{\nabla M}{M} = -\tan\sigma\begin{bmatrix}1\\0\end{bmatrix} \quad (9)$$

We note that

$$m_2 = m_1 + \nabla m \circ (\Delta t \quad \Delta b)^T$$
$$M_2 = M_1 + \nabla M \circ (\Delta t \quad \Delta b)^T$$

to first order where $\mathbf{p_2} - \mathbf{p_1} = (\Delta t \quad \Delta b)^T$ is the step between the points $\mathbf{p_1}$ and $\mathbf{p_2}$ in the image. Using this equation and equation 8, we get

$$A = \mathrm{Rot}(\delta_t) \cdot \begin{bmatrix} k_m \cos\delta_T & k_m \sin\delta_T \cos\sigma \\ -k_M \frac{\sin\delta_T}{\cos\sigma} & k_M \cos\delta_T \end{bmatrix}$$

where

$$k_m = 1 + \frac{\nabla m}{m} \circ \begin{bmatrix} \Delta t \\ \Delta b \end{bmatrix}, \quad k_M = 1 + \frac{\nabla M}{M} \circ \begin{bmatrix} \Delta t \\ \Delta b \end{bmatrix} \tag{10}$$

The actual affine transformation we find between the two points will not, in general, be in terms of the $(\mathbf{t}, \mathbf{b})$ basis. Instead, our matrix will be related by a change of basis: $\hat{A} = UAU^{-1}$. The change of basis matrix, U, rotates the standard basis to the $(\mathbf{t}, \mathbf{b})$ basis and is given by $\mathrm{Rot}(\theta_t)$, where θ_t is the tilt angle at point $\mathbf{p_1}$.

The analysis above assumes that we have spherical projection. In reality, cameras use planar projection. We can convert between the two types of projection by applying the Jacobian of the gaze transformation [7]. For each of the examples in this paper the gaze transformation was insignificant, so we ignored it.

4 Shape Recovery Algorithm and Experimental Results

In this section, we develop a new algorithm for recovering surface orientation (slant and tilt) and shape (principal curvatures and directions), with an associated sensitivity analysis. We will also present experimental results on a number of images of planar and curved objects, and discuss predictions for human perception of shape from texture.

There are five unknowns: the slant σ, the tilt direction specified by θ_t, and the three shape parameters $(\tau\kappa_t, \tau\kappa_b, \tau\tau)$. Each estimation of an affine transform in an image direction $\mathbf{v_i} = (\Delta t_i \quad \Delta b_i)^T$ yields four nonlinear equations. Simple equation counting tells us that one direction is not enough, and generically two directions ought to be sufficient.

Our shape recovery algorithm is based on finding the orientation and shape parameters which minimize the sum of squared errors between the predicted and empirically measured entries of the affine transformation matrices, i.e., we wish to minimize the following error function:

$$\chi^2(\sigma, \theta_t, \tau\kappa_t, \tau\kappa_b, \tau\tau) = \sum_{i=1}^{n} \sum_{k=1}^{2} \sum_{l=1}^{2} (\hat{A}_i(k,l) - \bar{A}_i(k,l))^2$$

where $\hat{A}_i(k,l)$ the (k,l)th element of the theoretically predicted matrix $\hat{A}_i$ and is a function of the shape parameters, and $\bar{A}_i(k,l)$ is the (k,l)th element of the empirically measured affine transform matrix $\bar{A}_i$. Ideally, each term in this error sum should be weighted by the inverse of the standard deviation of the measurement error of that particular entry. A specific characterization of the probability

Table 1. True and Estimated Surface Parameters

	True					*Estimated*					
Image	σ	tilt	$\tau\kappa_t$	$\tau\kappa_b$	$\tau\tau$	σ	tilt	$\tau\kappa_t$	$\tau\kappa_b$	$\tau\tau$	λ^2_{min}
cane	69	-90	0	0	0	71	-92	-.06	.00	-.03	.031
wire	64	-25	0	0	0	63	-20	.22	.00	-.03	027
noise	64	-25	0	0	0	61	-18	.32	-.01	-.10	.024
straw	64	-90	0	0	0	60	-73	-.13	.29	-.27	.010
cyl	28	180	6.5	0	0	26	180	3.2	-.13	-.15	.003
sph1	39	180	6.6	6.6	0	40	179	6.0	5.8	.67	.002
sph2	40	180	6.6	6.6	0	35	182	8.0	5.6	.95	002

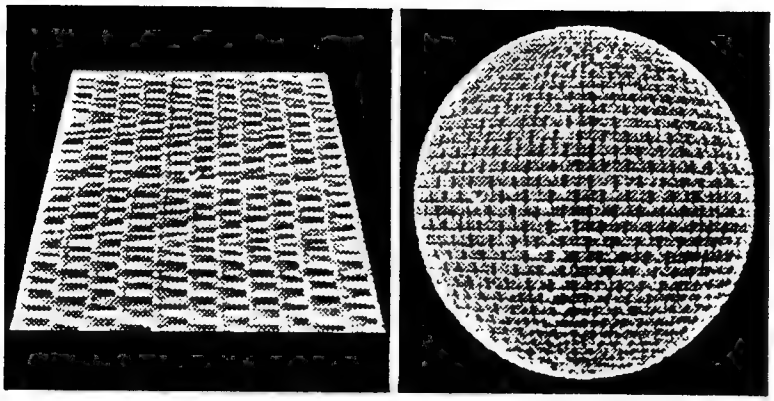
distribution of the measurement errors in the entries of the affine transform matrices A_i is not yet available. We expect it to depend on the particular algorithm used for the estimation of the affine transforms. In the absence of a particularly appropriate model, we will proceed on the (convenient!) assumption that these errors are independent and normally distributed with standard deviation Δa.

For minimizing the error function, we just used the gradient descent routine in the *Mathematica* package–there are any of a number of variants, such as conjugate gradient, that could have been used equivalently We obtain an initial guess for the shape parameters $(\sigma, \theta_t, \tau\kappa_t, \tau\tau)$ using a linear algorithm based on the singular value decomposition of the $\bar{A}_i$ matrices [13]. We initially set $\tau\kappa_b$ equal to the initial value of $\tau\kappa_t$.

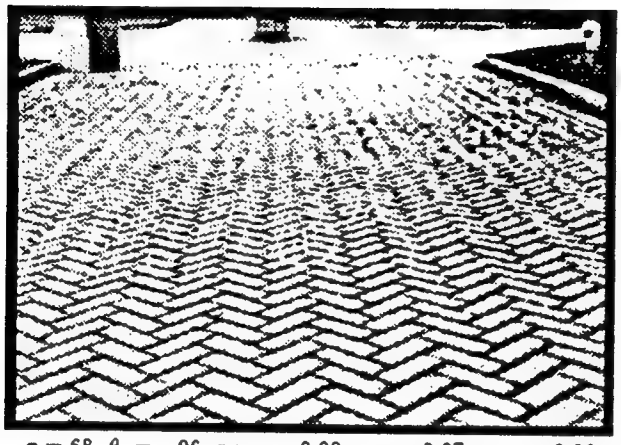
Table 1 shows the results on a number of synthetic examples. The image "noise" is the "wire" image, with added noise of standard deviation 30. The first four surfaces are planar, followed by a cylinder and two spheres. We also ran our algorithm on two real images. Figure 4 shows two of the synthetic images, as well as two natural images. We get better slant and tilt estimates than that provided by our simple linear algorithm, and also significantly better estimates of the curvature parameters [13]. The algorithm does quite well even in the presence of noise. We get good results on the "straw" planar surface, even though this texture does not satisfy the isotropy assumption commonly used by other researchers. Since we do not have ground truth for the natural images, we indicate the computed surface orientation by a projected circle in the image, and give the shape estimates in the captions. We get quite reasonable orientation estimates, and believable curvature estimates, for both of the natural images.

We will now discuss the determination of confidence intervals on the orientation and shape estimates. For more details on the methodology that we follow, the reader is referred to [16], Chapters 14.4 and 14.5.

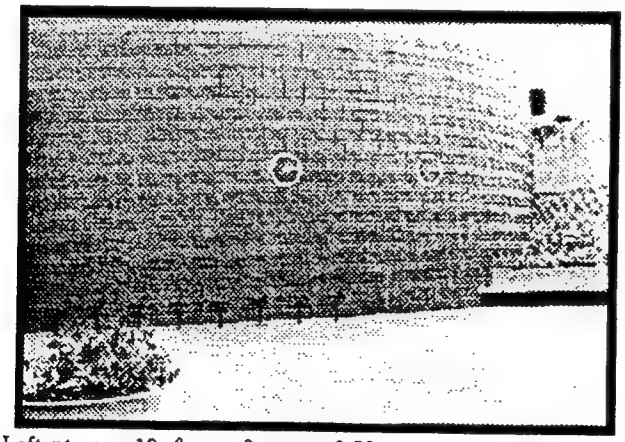
Let us abbreviate the five geometrical parameters $(\sigma, \theta_t, \tau\kappa_t, \tau\kappa_b, \tau\tau)$ as $g_i, i = 1, 2, \ldots 5$. The gradient of χ^2 with respect to the parameters g will be zero at the χ^2 minimum. To obtain confidence intervals on the parameters, one computes the so-called ***curvature matrix*** $[\alpha]$ which is defined as half of the Hessian of the



straw sph1



$\sigma = 68$, $\theta_t = -96$, $\tau\kappa_t = -0.08$, $\tau\kappa_b = 0.07$, $\tau\tau = -0.06$



Left pt: $\sigma = 19$, $\theta_t = -8$, $\tau\kappa_t = 0.53$, $\tau\kappa_b = -0.12$, $\tau\tau = 0.00$
Right pt: $\sigma = 40$, $\theta_t = -11$, $\tau\kappa_t = 2.9$, $\tau\kappa_b = -0.03$, $\tau\tau = 0.58$

Fig. 3. Experimental images.

χ^2 function

$$\alpha_{kl} = \frac{1}{2\Delta a^2} \frac{\partial^2 \chi^2}{\partial g_k \partial g_l}$$

Since we have five geometrical shape parameters g_i, this is a 5×5 symmetric matrix. The inverse of the curvature matrix is the covariance matrix, C, of the fit, on the assumption of normally distributed errors. In that case the standard errors in the parameters are given by $\sqrt{C_{ii}}$. The confidence interval for parameter g_i, $\pm\delta g_i$ is $\pm\sqrt{C_{ii}}$ for 68 percent confidence, $\pm 2\sqrt{C_{ii}}$ for 95 percent confidence.

As mentioned above, if we assume that the errors in the entries of the affine transformation matrices A_i are independent and identically normally distributed, and if we have an estimate for the standard deviation of these errors, we can give confidence intervals for the estimates of the shape parameters. We have calculated the 68 percent confidence intervals for the parameters, assuming[1] a measurement error Δa of standard deviation 0.0323. Using this value for the standard deviation, 66 percent of the estimated parameters for the synthetic examples fall within the 68 percent confidence intervals.

In addition to obtaining confidence intervals for our empirical shape estimates, we can use the theory outlined above to develop an ideal observer model for shape from texture. Roughly speaking, an ideal observer gives us a prediction for the best performance one expects out of any estimator, given the visual information and the measurement error. As such it is of interest both for computer vision shape from texture algorithms and for predicting the performance of the human visual system. In the context of shape from texture models based on discrete texels using isotropy and first order homogeneity assumptions, such ideal observers have been developed by Blake et al[3].

The uncertainty in the shape estimates depends upon the measurement errors in the earlier stages of processing; here, in the estimation of the affine transforms. The errors in the affine transforms will of course depend on the texture being viewed. For example, if the texture is a realization of a Poisson process then we expect that for sparser textures it will be more difficult to estimate the affine transforms accurately. Without prior knowledge of the distribution of the texture, however, what can we say about the distribution of errors in the affine transforms? A simple choice would be to assume, as we did previously, that the errors in the affine transform parameters are independent and identically distributed normal random variables. While clearly this assumption would be incorrect for, e.g., highly anisotropic texture, it gave us very reasonable results for the range of textures earlier in this section. One may think of the situation as follows: we have a number of surfaces with different shapes and orientations, all with similar textures which satisfy the above assumption, and we wish to know whether shape estimation is more difficult for some of these shapes than others. As above, we will find confidence intervals for the shape parameters for a

[1] We computed this value as the standard deviation of the errors in the affine transformation matrices for the examples used in this paper. The errors are known because we know the ground truth in these synthetic examples.

number of these hypothetical shapes. The confidence intervals give us a measure of the uncertainty in the shape estimates.

We present here the results of two sample ideal observer "experiments." For further "experiments," see [13]. Since the confidence intervals will depend on the measurement error in the elements of the affine transformations, we give the confidence intervals in terms of *relative units*. To obtain the actual expected error one would multiply these values by the standard deviation of the measurement error. In the first experiment in Fig. 4a, we varied the slant of a planar surface, and plot the confidence intervals for tilt estimates. We see that we expect improved tilt estimates for higher slants. Blake, et al[3] report similar results for their ideal observer based on compression gradient. In the second experiment, we varied the slant of a cylindrical surface, keeping the tilt perpendicular to the axis of the cylinder. This amounts to computing shape estimates for a series of points along the circumference of the cylinder. In Fig. 4b we plot the confidence intervals for the estimate of $r\kappa_t$. Note that for smaller slants we expect a more error in the curvature parameter. This may explain much of the error in $r\kappa_t$ for both our synthetic cylinder example and for the first point in the real cylinder example. Observations such as these suggest several lines of psychophysical investigation.

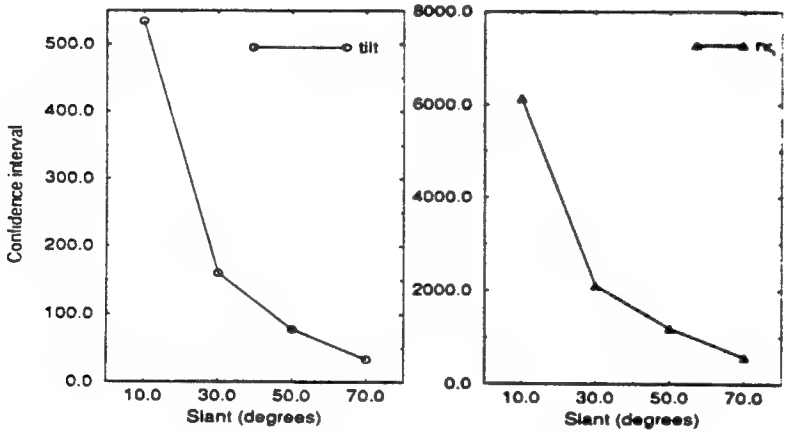


Fig. 4. Ideal observer experiments, described in text.

This research was supported by NSF PYI grant IRI-8957274, Xerox, the PATH project, and Joint Services Electronics Program contract no. F49620–93–C–0014. We thank Pietro Perona and Phil Lapsley for useful discussions, and John Oliensis for catching an error in an earlier version.

References

1. J. Aloimonos. "Shape from texture." *Biological Cybernetics*, Vol. 58, pp. 345-360, 1988.
2. R. Bajcsy and L. Lieberman. "Texture gradient as a depth cue." *CGIP*, 5:52-67, 1976.
3. A. Blake, H. Bulthoff, and D. Sheinberg, "Shape from texture: ideal observers and human psychophysics," *Vision Research*, 33(12):1723-37, Aug. 1993.
4. A. Blake and C. Marinos, "Shape from texture: estimation, isotropy and moments." *Artificial Intelligence*, 45(1990):323-380.
5. L.G. Brown and H. Shvaytser,"Surface orientation from projective foreshortening of isotropic texture autocorrelation." *IEEE Trans. on PAMI*, 12(6), June 1990, pp. 584-588.
6. L.S. Davis, L. Janos, and S.M. Dunn, "Efficient recovery of shape from texture." *IEEE Trans. on PAMI*, 5(5), 1983.
7. J. Gårding. "Shape from surface markings." Ph D. Dissertation, Dept. of Numerical Analysis and Computing Science, Royal Institute of Technology, 1991.
8. J. Gårding. "Shape from texture for smooth curved surfaces in perspective projection." *Journal of Mathematical Imaging and Vision*, 2(4):327-50, Dec. 1992.
9. K. Ikeuchi. "Shape From Regular Patterns." *J. of Artificial Intelligence*, 22:49-75, 1984.
10. K. Kanatani and T.C. Chou. "Shape from texture: General Principle." J. of Artificial Intelligence, Vol. 38, pp. 1-48, 1989.
11. J. Krumm and S. Shafer. "Shape from Periodic Texture Using the Spectrogram." Proc. CVPR, Champaign-Urbana, Illinois, 1992, 284-301.
12. J.Malik and R. Rosenholtz. "A Differential Method for Computing Local Shape-From-Texture for Planar and Curved Surfaces." Proc. CVPR, New York, 1993.
13. J. Malik and R. Rosenholtz, "A Differential Method for Computing Local Shape-From-Texture for Planar and Curved Surfaces," CS Division Technical Report, UCB-CSD-93-775, UC Berkeley, 1993.
14. C. Marinos and A. Blake. "Shape from texture: the homogeneity hypothesis." Proc. ICCV, Osaka, Japan, 1990, pp.350-353.
15. B. O'Neill. "Elementary Differential Geometry." New York: Academic Press, 1966.
16. W.H. Press, B.P. Flannery, S.A. Teukolski, W.T. Vetterling. "Numerical Recipes in C." Cambridge University Press, 1988.
17. B. Super and A. Bovik. "Shape-from-Texture by Wavelet-Based Measurement of Local Spectral Moments." Proc. CVPR, Champaign-Urbana, Illinois, 1992, 296-301.
18. A.P. Witkin. "Recovering surface shape and orientation from texture." J. of Artificial Intelligence, Vol. 17, pp. 17-45, 1981.

Characterization of Hot-Carrier Effects in Thin-Film Fully-Depleted SOI MOSFETs

Z.J. Ma, H.J. Wann, M. Chan, J. King,
Y.C. Cheng*, P.K. Ko, and C. Hu

Dept. of EECS, University of California, Berkeley, CA 94720, U.S.A
*The Directorate, City Polytechnic of Hong Kong, Hong Kong

ABSTRACT

Previous conflicting reports concerning fully-depleted SOI device hot electron reliability is partially due to misunderstanding over the maximum channel electric field (E_m). Experimental results using SOI MOSFETs with body contacts indicate that E_m is just a weak function of thin-film SOI thickness (T_{si}) and E_m can be significantly lower than in a bulk device with drain junction depth (X_j) comparable to T_{si}. The theoretical correlation between SOI MOSFETs gate current and substrate current are experimentally confirmed. This provides a means (I_G) of studying E_m in SOI device without body contacts. Both N- and P-MOSFETs can have better hot-carrier reliability than comparable bulk devices. Thin film SOI MOSFETs have better prospects for meeting breakdown voltage and hot-electron reliability requirements than previously thought.

I. INTRODUCTION

Thin-film fully-depleted (FD) SOI MOSFETs have attracted much attention because of their large drain saturation current, absence of kink effect, and superior subthreshold leakage. However, as far as device reliability and breakdown voltage are concerned, previous reports are divided on whether FD SOI devices have reduced or enhanced hot-carrier susceptibility and sensitivity to SOI film thickness [1]-[7]. The main difficulty is that substrate current, (I_{SUB}), the convenient monitor of channel field for bulk MOS devices, cannot be measured on the SOI devices with floating body. While gate current has been suggested as a lifetime parameter for SOI N-MOS devices [6], gate current is difficult to measure and gate current has not been confirmed as a valid monitor of channel field without substrate current [7]. In this study, for the first time, using a special SOI device structure with body contact, both substrate (body) current (I_{SUB}) and gate current (I_G) were directly measured in the same devices. Based on this experiment, not only is the channel field in SOI devices quantified, but also the correlation between I_{SUB} and I_G is established. Finally, the hot carrier degradation lifetimes of SOI N- and P-MOSFETs are compared with those of bulk devices.

II. EXPERIMENTS

FD N- and P-channel SOI devices were fabricated SIMOX wafers using a CMOS process. The buried o thickness is about 3500 Å. LOCOS isolation and 3000 Å in-doped N^+ poly gate were used. To collect I_{SUB}, the N-M devices have a special P^+ region to contact the P-type body illustrated in Fig.1. The P^+ region was formed by P-MOS implant (B_{11}, $4x10^{15}$ cm^{-2}, 30 keV). Similarly, P-M devices have an N^+ region to contact N-type body using N-M S/D implant (As, $4x10^{15}$ cm^{-2}, 70 keV). The measured I_{SUI} these structures is found to be proportional to the channel wi indicating a high efficiency in collecting I_{SUB}.

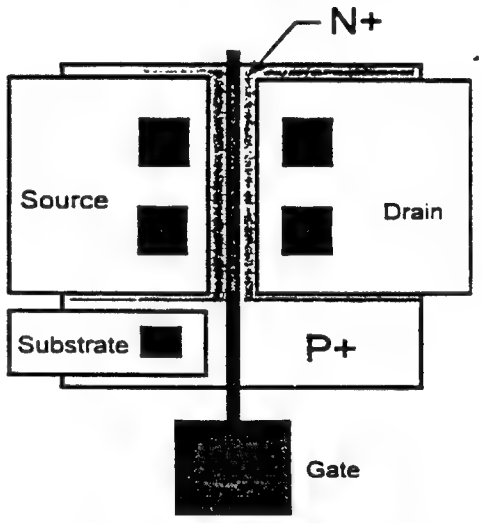


Fig.1 *A schematic of the body contact in N-MOS SC devices. SOI P-MOS devices have the similar structu with N^+ and P^+ regions exchanged.*

III. RESULTS AND DISCUSSIONS

(A) Analytical Hot-Carrier Models

In bulk MOSFETs, the maximum channel field can estimated as follows [8]:

CH3332-4/94/0000-0052$01.00 © 1994 IEEE/IRPS

$$E_m = (V_D - V_{DSAT})/\ell \quad (1)$$

$$\ell = 0.22 T_{OX}^{1/3} X_j^{1/2} \quad (2)$$

where X_j is the drain junction depth and ℓ is the characteristic length. The hot-carrier currents, I_{SUB} and I_G, can be expressed as follows [9]:

$$I_{SUB} = \frac{A_i}{B_i}(V_D - V_{DSAT}) I_D \exp(-\frac{B_i}{E_m}) \quad (3)$$

where A_i and B_i are known constants for impact ionization rate of channel carriers [9]. B_i is around 1.7×10^6 V • cm^{-1} for electrons and 3.7×10^6 V • cm^{-1} for holes [10].

$$I_G = C_1(E_{ox}) I_D \exp(-\frac{\varphi_b}{E_m \lambda_e}) \quad \text{for N-MOS,} \quad (4)$$

$$I_G = C_2(E_{ox}) I_{SUB} \exp(-\frac{\varphi_b}{E_m \lambda_e}) \quad \text{for P-MOS,} \quad (5)$$

where φ_b is the barrier height at Si/SiO_2 interface for electron, λ_e is the scattering mean-free path of electron. It is worth noting that the gate current of P-MOSFET's results from electron injection rather than hole injection into the oxide in the low-V_G regime [11].

From Eqs. 3, 4, and 5, we find:

$$\ln\left(\frac{I_{SUB}}{I_D(V_D - V_{DSAT})}\right) = \ln(\frac{A_i}{B_i}) - \frac{B_i \ell}{V_D - V_{DSAT}} \quad (6)$$

$$\frac{I_G}{I_D} = \frac{C(E_{ox}) B_i}{A_i (V_D - V_{DSAT})} (\frac{I_{SUB}}{I_D})^{\frac{\varphi_b}{B_i \lambda_e}} \quad \text{for N-MOS} \quad (7)$$

$$\frac{I_G}{I_{SUB}} = \frac{C(E_{ox}) B_i}{A_i (V_D - V_{DSAT})} (\frac{I_{SUB}}{I_D})^{\frac{\varphi_b}{B_i \lambda_e}} \quad \text{for P-MOS.} \quad (8)$$

(B) Hot-Carrier Currents and Channel Electric Field

In SOI devices, whether Eq. 1 and Eq. 2 can be used and how are not clear. Colinge [3] used T_{si} as a substitute for X_j in Eq.2 to estimate the channel field, while Chen, et al. [6] reported that this overestimates E_m. From Eq. 6, a plot of $\ln(I_{SUB}/(I_D(V_D - V_{DSAT})))$ versus $1/(V_D - V_{DSAT})$ should yield one straight line for all bias voltages, for both N- and P-MOSFETs, as shown in Fig.2. The slope of this straight line gives $B_i \ell$, from which ℓ and hence E_m can be determined experimentally. It is found that using thin film SOI thickness T_{si} as a substitute for X_j can overestimate E_m by a factor of 2 for both N- and P-MOS devices (roughly equivalent to overestimating V_D by more than 50%).

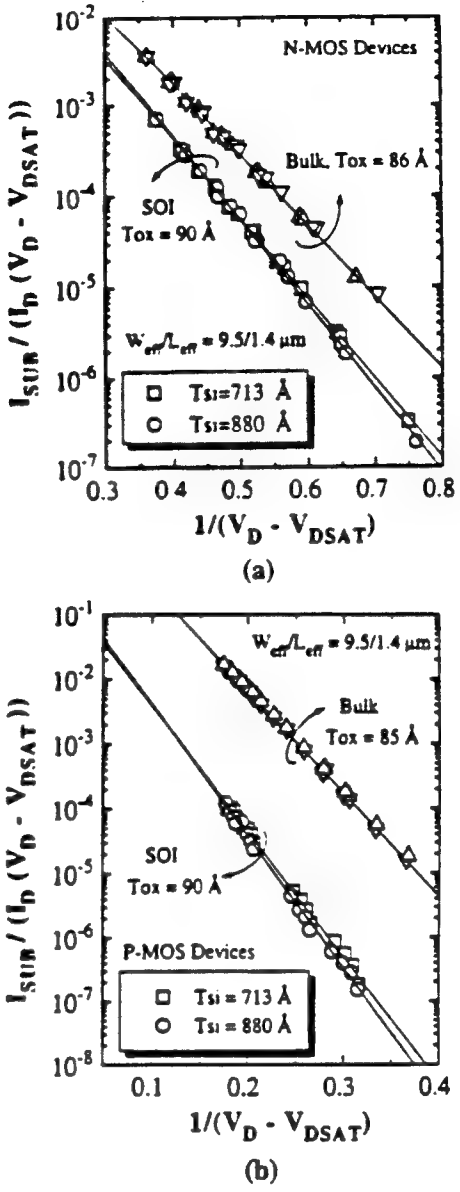
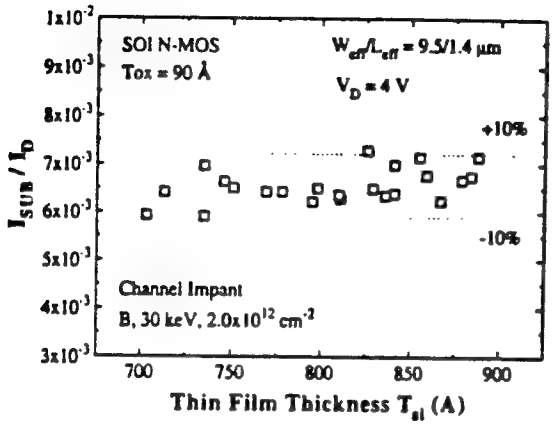


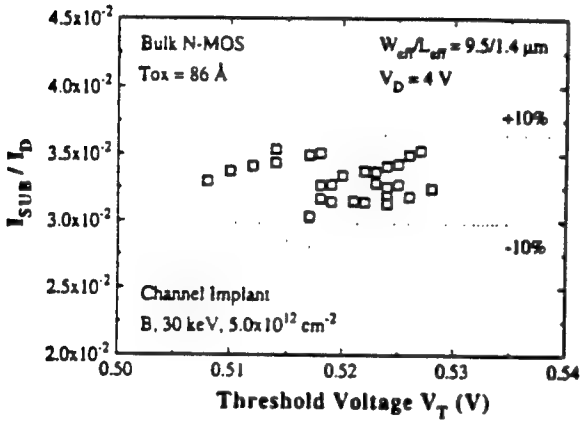
Fig.2 *Experimental determination of the characteristic length ℓ in $E_m = (V_D - V_{DSAT})/\ell$ in SOI and bulk devices. (a), N-MOS devices: For bulk device, the measured ℓ is 461 Å. For SOI device, the measured ℓ s are 1194 and 1254 Å, while the ℓ s, based on the bulk MOSFET E_m model with $X_j = T_{si}$, are 554 Å and 616 Å. (b), P-MOS devices: For bulk device, the measured ℓ is 426 Å. For SOI device, the measured ℓ s are 1213 and 1271 Å, while the ℓ s, based on the bulk MOSFET E_m model with $X_j = T_{si}$, are 554 Å and 616 Å. The data clearly show that E_m in SOI devices can be much lower than in bulk devices.*

It can be deduced from Eq.3 that I_{SUB}/I_D is a simple monitor of the maximum channel field E_m. Fig.3(a) and (b) show the distribution of measured I_{SUB}/I_D for the N-MOS SOI

and bulk devices, respectively. Although T_{si} variation across a wafer is as high as 200 Å, the I_{SUB}/I_D variation is only ±10% and comparable to the ±10% variation in the bulk case, confirming the weak E_m dependence on T_{si}. Besides, SOI devices have much lower I_{SUB}/I_D, hence E_m, by a factor of 4 than comparable bulk devices (see Fig3(a) and (b)), although V_{DSAT} are about the same. T_{si} was determined by the CV technique [12].



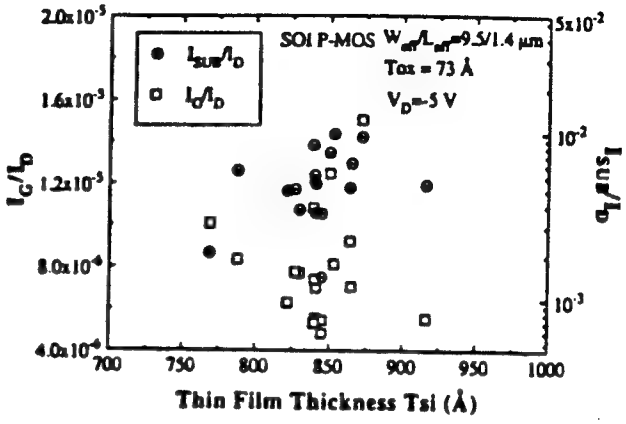
(a)



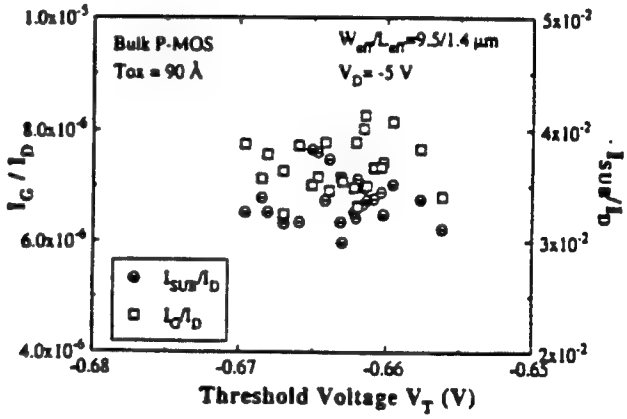
(b)

Fig.3 *Distribution of I_{SUB}/I_D of (a) SOI , and (b) bulk N-MOS devices. Both I_{SUB} and I_D were measured at the maximum I_{SUB} point with V_D=4 V. The I_{SUB}/I_D sensitivity to SOI T_{si} is very weak. The variations of I_{SUB}/I_D across both wafers are about ±10%. SOI devices have lower I_{SUB}/I_D hence E_m by a factor of 4 than the comparable bulk MOSFETs.*

The P-MOS data is shown in Fig.4. It can be seen that the measured I_{SUB}/I_D is also a weak function of T_{si}. The I_{SUB}/I_D values for the SOI P-MOS devices are lower than that in the bulk devices even though the SOI devices have a thinner gate oxide. Since the gate current in P-MOS devices is much larger as compared to that in N-MOS devices thus easier to be measured, the I_G/I_D data is also presented in the same figure. Clearly, the electron injection into the gate oxide is also a weak function of the T_{si}. Besides, the I_G/I_D values for the SOI devices are a little higher than the bulk devices probably because the vertical field of gate oxide on the SOI devices is higher than on the bulk devices.



(a)



(b)

Fig.4 *Distribution of I_{SUB}/I_D and I_G/I_D of (a) SOI, and (b) bulk P-MOS devices. I_{SUB}, I_G and I_D were measured at the maximum I_{SUB} point with V_D=-5 V. The I_{SUB}/I_D and I_G/I_D sensitivity to SOI T_{si} is very weak. SOI devices have lower I_{SUB}/I_D and hence E_m than the bulk MOSFETs.*

Fig.5 demonstrates the combined effects of 2D and lateral doping gradient in drain region. The channel field increases exponentially due to the 2D effect, while decreases near the drain region due to the lateral doping effect. Both the lower E_m and weaker dependence on T_{si} in thin film SOI devices can be attributed to the lateral drain doping gradient [13]. Simulations found relatively low E_m and weak E_m sensitivity on T_{si} within the range of interest (500~1100 Å). E_m is a function of the lateral drain doping gradient even for non-LDD As drains. In

the case of bulk MOSFETs, the doping gradient varies with X_j; this contributes to the E_m dependence on X_j. In the case of SOI MOSFETs, lateral doping gradient is decoupled from X_j (or T_{si}). Therefore, E_m can be lower in an SOI device than a bulk device with small $X_j \approx T_{si}$ and a weak dependence on T_{si} as well.

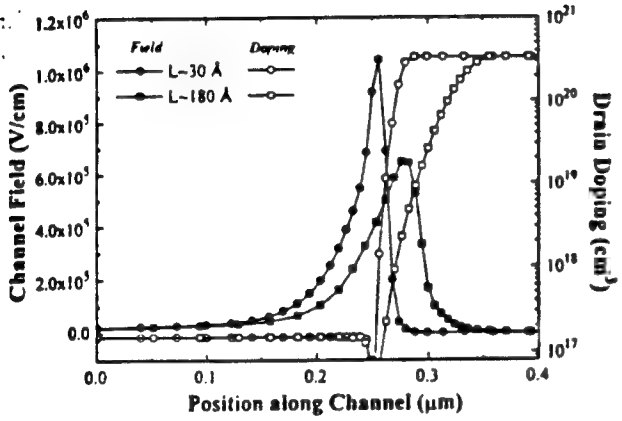


Fig.5 *Demonstration of the 2D and lateral drain doping gradient effects on channel electric distributions in a MOS device. The lateral doping gradient L is defined as*

$$L = \left(\frac{1}{N_D} \frac{dN_D}{dy} \right)^{-1} \text{ at around } N_D = 1\text{-}5 \times 10^{18}\ cm^{-3}.$$

(C) Hot-Carrier Effects

I_{SUB} and I_G are correlated to each other by a power-law relationship, because both of them are exponential functions of E_m in equations 3, 4 and 7 [14],[15]. Fig.6 demonstrates the relationship between I_{SUB} and I_G in SOI N-MOS devices. Based on equation 7, to the first order, the $\log(I_G/I_D)$ versus $\log(I_{SUB}/I_D)$ plot is a function of only oxide field, $E_{ox}=(V_G-V_D)/T_{ox}$, and its slope is equal to $\varphi_b / (B_i\lambda)$. This slope is 2.1 suggesting that $B_i\lambda$, i.e., the critical electron energy for impact ionization, is about 1.5 eV if we assume the Si/SiO_2 barrier height for electrons is φ_b=3.1 eV. This result agrees very well with the bulk case [15]. The agreement between theoretical prediction and experimental data suggests that I_G can be used as a monitor of E_m for SOI N-MOS devices with the body contacts, e.g., ℓ can be estimated from the slope of the $\log(I_G/I_D)$ versus $1/(V_D-V_{DSAT})$ plot.

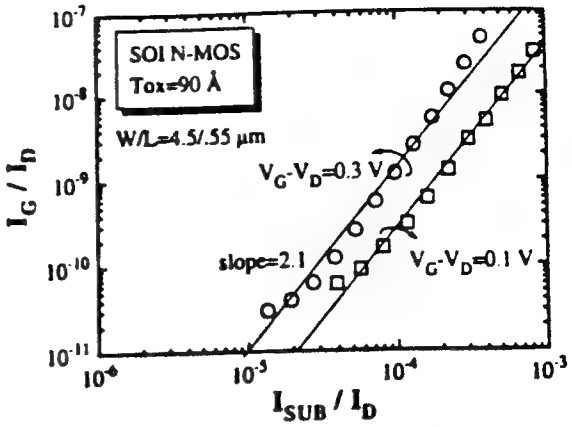


Fig.6 *Correlation between I_{SUB} and I_G in SOI N-MOS devices. The slope is about 2.1. Based on the expression (4), the slope approximately equals to $\varphi_b / (B_i\lambda)$ from which $B_i\lambda$ can be calculated.*

Previous reports are divided on whether FD SOI devices are less [3]-[7] or more [1], [2] vulnerable to hot-carrier effects. The fact that FD SOI devices often have a lower channel field E_m than bulk devices due to a more graded drain doping profile should also be reflected from the device degradation in terms of hot-carrier stress. In addition, Fig.7 shows the saturation drain current shifts for a given maximum I_{SUB} is not larger in SOI than bulk N-MOS devices. Fig.8 shows the extrapolated lifetime for both SOI N- and P-MOS devices, respectively, as compared to the bulk devices. The devices were stressed under the worse-case stress conditions (maximum I_{SUB} for N-MOS case and maximum I_G for P-MOS case). Although X_j for the bulk devices is as large as 2000 Å, the SOI devices with T_{si}=800 Å are still slightly less vulnerable than the bulk devices to hot-carrier degradation at least in this case study.

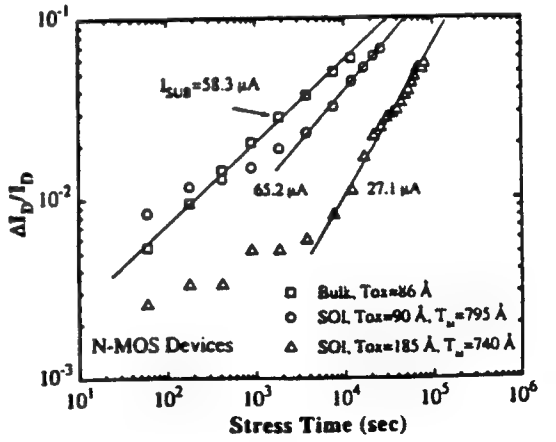


Fig.7 *Comparison of the saturation drain current shifts for a specific maximum I_{SUB} stress between the SOI and bulk N-MOS devices.*

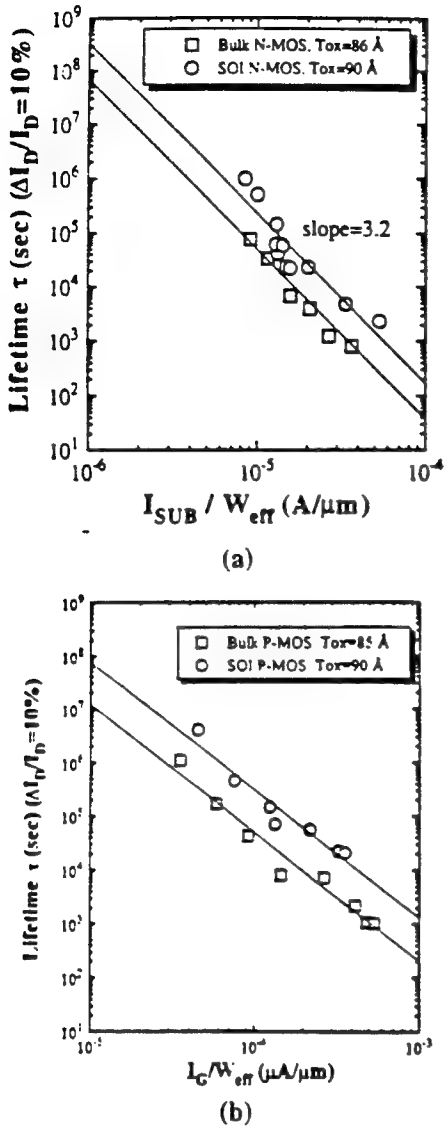


Fig.8 *Device hot-carrier lifetime of the SOI and bulk devices as a function of, (a), substrate current in N-MOS devices; and, (b) gate current in P-MOS devices. The lifetime τ is defined as the stress time to reach 10% $\Delta I_D/I_D$.*

IV. CONCLUSIONS

In conclusion, channel field in SOI MOSFETs is lower than previously assumed. Hot-carrier degradation of thin film FD SOI devices can be less severe than a similar bulk MOSFET. The decoupling of SOI MOSFET junction depth (T_{si}) and lateral doping gradient is little discussed but of significant advantage in drain engineering. This realization improves the prospects of thin film SOI devices meeting breakdown voltage and hot-carrier effects requirements. The correlation between I_{SUB} and I_G is confirmed and suggests that the channel field may be characterized through the measurable I_G. I_G measurement should help drain structure design engineering.

ACKNOWLEDGMENT

This work was supported by SRC, TI, Rockwell International under California state MICRO program, and AFOSR/JESP under contract F49620-90-0029.

REFERENCES

[1] P.H. Woerlee, A.H. Ommen, H. Lifka, C.A.H. Juffermans, L. Plaja, and F.M. Klaassen, "Half-micron CMOS on ultra-thin silicon on insulator," IEDM Tech. Dig., p.821, 1989.

[2] P.H. Woerlee, C. Juffermans, H. Lifka, W. Manders, F.M.O. Lansink, G.M. Paulzen, P. Sheridan, and A. Walker, "A half-micron CMOS technology using ultra-thin silicon on insulator," IEDM Tech. Dig., p.583, 1990.

[3] J.P. Colinge, "Hot-carrier effects in silicon-on-insulator n-channel MOSFET's," IEEE Trans. Electron Devices, vol.10, p.2173, 1987.

[4] J.G. Fossum, J.Y. Choi, and R. Sundaresan, "SOI design for competitive CMOS VLSI," IEEE Trans. Electron Devices, vol.37, p.724, 1990.

[5] S. Cristoloveanu, S.M. Gulwadi, D.E. Ioannou, G.J. Campisi, and H.L. Hughes, "Hot-electron-induced degradation of front and back channels in partially and fully depleted SIMOX MOSFET's," IEEE Electron Device Lett., vol 13, p.603, 1992.

[6] J. Chen, K. Quader, R. Solomon, T.Y. Chan, P.K. Ko, and C. Hu, "Hot electron gate current and degradation in p-channel SOI MOSFET's," IEEE SOS/SOI Conference, p.8, 1991.

[7] L.T. Su, H. Fang, J.E. Chung, and D.A. Antoniadis, "Hot-carrier effects in fully-depleted SOI NMOSFETs," IEDM Tech. Dig., p.349, 1992.

[8] T.Y. Chan, P.K. Ko, and C. Hu, "Dependence of channel electric field on device scaling," IEEE Electron Device Lett., vol.EDL-6, p.551, 1985.

[9] P.K. Ko, "Hot-electron effects in MOSFETs," Ph.D thesis, University of California at Berkeley, 1982.

[10] T.C. Ong, P.K. Ko, and C. Hu, "Modeling of substrate current in p-MOSFET's," IEEE Electron Device Lett., vol.EDL-8, p.413, 1987.

[11] T.C. Ong, K. Seki P.K. Ko, and C. Hu, "P-MOSFET gate current and device degradation," in Porc. of IEEE IRPS, p.178, 1989.

[12] J. Chen, R. Slomom, T.Y. Chan, P.K. Ko, and C. Hu, "A CV technique for measuring thin SOI film thickness," IEEE Electron Device Lett., p.453, 1991.

[13] H.J. Wann, J. King, Jian Chen, P.K. Ko, and C. Hu, "Hot-carrier currents of SOI MOSFETs," IEEE SOS/SOI Conference, p.118, 1993.

[14] S.Tam, P.K. Ko, C. Hu, and R.S. Muller, "Correlation between substrate and gate currents in MOSFTE's," IEEE Trans. Electron Devices, p.1740, 1982.

[15] C. Hu, "Hot-electron effects in MOSFET's," IEDM Tech. Dig., p.176, 1983.

Current Research in Acoustically Robust Speech Recognition

Nelson Morgan
International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704, USA.
(510)-643-9153

July 13, 1994

Current Research in Acoustically Robust Speech Recognition

Nelson Morgan
International Computer Science Institute
and UC Berkeley

Abstract

While recognizer technologies are being applied to increasingly challenging problems, it is still true that the robustness of these systems to acoustic variability (such as room acoustics, background noise, and modified microphone or channel characteristics) still is far poorer than that achieved by human listeners. A wide variety of techniques have been developed by researchers to attempt to deal with these problems. This paper will describe the current state of our approach to this problem at Berkeley, which has focused on the modeling of some gross temporal properties of human hearing.

1 INTRODUCTION

Speech processing, and in particular speech recognition by machine, is severely hampered by problems such as the effect of room acoustics, background noise, and mismatches between recording characteristics for the systems used for training and recognition (for instance, different handsets for the telephone).

Human beings are also hampered in their speech understanding by these factors, but we appear to do much better. Certainly much of this robustness is due to our ability to predict and interpret speech on the basis of our understanding of the language and our knowledge of the world and how it works. However, even the recognition of digits in the presence of strong additive noise is quite a bit better for people than for our best recognizers. This is a task that uses relatively little knowledge about language other than knowing the digits themselves, since in many applications any digit can follow any other. Therefore, we think that it is reasonable to try to learn acoustical processing approaches from human solutions to these problems in order to guide our study of the design of robust machine recognizers.

The oldest engineering methods for improving speech processing in the presence of noise and spectral coloration are based on Wiener filtering. The basic idea of this approach is to estimate the noise spectrum and the speech spectrum and find a good compromise between damaging the speech and getting rid of the noise. For instance, in a simple case in which the noise and speech were in different parts of the audio spectrum, you could simply filter out the noise. In most practical cases, the problem is not this easy. However, if the noise doesn't change too much over time, is uncorrelated with the speech, and its spectrum can be estimated accurately, one can also subtract out much of its effects. In the case of additive noise, this approach is commonly called spectral subtraction [1]. In the case of linear filtering effects (for instance the effect that results from using for training and testing telephone channels that have differing spectral characteristics), this subtraction is done in the logarithmic spectral domain and is closely related to what has been called

blind deconvolution. This was named for the process used to remove much of the "tinny" audio characteristics on old Caruso recordings by normalizing out the linear effects of the mechanical gramaphone recording horn [8]. In more recent years similar techniques have been used to normalize out the effect of the frequency responses of different microphones: this is commonly done in the *cepstral* domain, which is a Fourier decomposition of the log spectrum. The technique is commonly called *mean cepstral subtraction*, since it basically consists of subtracting the average cepstrum (or equivalently the average log spectrum) from the speech.

These methods have helped to some extent, but it is still true that adding even a moderate amount of realistic noise has a strongly negative effect on speech recognition systems. For this reason we have focused on developing algorithms that have better robustness to unpredictable acoustic test conditions. We have tended to use methods inspired by human hearing, although we depart from the biological example whenever it seems to make sense. This could be likened to the development of airplanes that utilize Bernoulli's principle, which correctly describes the physics that permits birds to fly, without blindly following the exact design and trying to build 747's that flap their wings. In our case, much of our work has been based on the observation that the human nervous system is most responsive to novelty, and in general that there is a certain range of speeds of sensory phenomena that we are most sensitive to. This is a fundamental fact about our sensory systems, and human speech and language developed in this context; as a species we developed speech in a form that we could hear well. This suggests that our machine speech analysis systems might benefit from these timing-based factors. Over the last few years, we have performed a number of experiments that suggest that this is a worthwhile strategy. The rest of this paper will briefly describe an approach we have developed to take advantage of one of these properties of human hearing, and an experiment that indicates the utility of the approach on an isolated digits task. The paper will conclude with a discussion of the future directions in our speech robustness work at ICSI and UC Berkeley (in collaboration with the Oregon Graduate Institute).

2 RASTA PROCESSING

As mentioned above, natural sensory systems tend to respond more strongly to novel stimuli rather than to continuations of the same old thing. This strategy has some obvious consequences for survival of the organism (you would really want to know about the rustle in the bushes, as it might be due to something that wants to eat you). Additionally, however, sensitivity to novelty tends to have a normalizing property. If you are most interested in change, then a constant background (for instance, of illumination, color, or acoustic noise) will have less of an effect on perception. This is desirable, since the message is often the same regardless of the background conditions.

Many experiments have been done with human listeners to determine the sensitivity to different aspects of speech. In one such experiment, for instance, Summerfield and colleagues showed that the perception of speech-like sounds depends on the spectral difference

between the current and preceding sounds [9]. Other experiments (for instance those reported in [4]) showed that human listeners were relatively insensitive to slowly varying sounds. This may partially explain why human listeners do not seem to mind a slow change in the frequency characteristics of the communication environment, or why steady background noise often does not severely impair human speech communication.

Thus, to make speech analysis less sensitive to slowly changing or steady-state factors in speech, we developed a method based on a simple bandpass or highpass filtering of the time trajectories of each spectral channel. This is equivalent to a spectrum based on change, and so we called it the RelAtive SpecTrAl or RASTA method [3]. Initially our experiments were only done by filtering in the log spectral domain, which is the optimal function to filter for the case in which some relatively constant linear filtering had been done on the speech (for instance, to represent the frequency response of the microphone or hand set). These experiments showed a greatly increased robustness to strong channel variation, and variants of the methods have been used by many laboratories worldwide. However, the method initially showed no increased robustness to additive noise.

More recently, we developed an extension to the original RASTA called J-RASTA (or sometimes lin-log RASTA) [6, 5]. In this method, the domain in which the speech spectrum was filtered or "relative" was a function of the noisiness of the data. For very noisy data, the filtering was done on something very close to the power spectrum, and for clean data the filtering was done on something closer to the log power spectrum. Typically we used a single family of functions that was parameterized by a "J" variable that was inversely proportional to the estimate of noise power derived from the local statistics [2].

This technique has recently been implemented in public domain software, and was shown to be effective against both linear filtering and additive noise. Table I shows some results on an isolated digits experiment. For this experiment, 200 speakers were used; in each of 4 tests, 150 speakers were used for training and 50 for test. By rotating (or "jackknifing") which speakers were the training or test portion, all of the data could be fairly used for testing. The HMM Tool Kit (HTK) from Cambridge [10] was used to design and train a Gaussian-mixture based speech recognizer.

Table I shows that the RASTA approach, when applied in the J-RASTA formulation, appears to provide some robustness to the effects of additive noise and linear filtering. In particular, the error rate with added noise was one-third of that seen without the RASTA processing. However, adding noise still quadrupled the error rate, even at a 10 dB SNR. At this signal-to-noise ratio most humans would have very little increase in error on this task. Therefore it is fair to say that more work needs to be done, and perhaps incorporating some other simple characteristics of human hearing may be helpful.

A number of researchers, including some at Berkeley, are working to employ better auditory models at the front end of speech recognition systems. This may in fact prove important. However, for the most part we are focusing on the importance of changing the overall statistical system in order to accommodate such representations. This is briefly described next.

	clean	noise	clean-filtered	noise-filtered
PLP	5.0	37.0	24.9	50.4
RASTA	3.3	50.0	3.6	40.4
J-RASTA	3.7	13.7	5.6	17.1

Table 1: Isolated digit (plus the words "yes" and "no") error rates in percent, using HTK-based Gaussian mixture system. PLP was the basic speech feature set used, which was modified for robustness in the other analysis methods. For all cases, the recognizer was trained on clean speech. Noise indicates test data with SNR=10dB, and noise-filtered indicates test data with the SNR=10dB and with an additional linear distortion introduced by filtering. Note that all of the error rates are somewhat high for an isolated digits recognizer; one of our recognizers with more features and a more advanced statistical model has roughly one-third the error rate, but currently takes much longer to train and so was not used for this study.

3 PERCEPTUALLY-BASED STATISTICAL MODELS

Traditionally, pattern recognition systems such as speech or character recognizers have been divided into two major components: feature extraction (including all kinds of preprocessing) and pattern matching. However, it is well known that the two components are strongly interdependent. For instance, for the case of good acoustic conditions (matching in training and test) we have sometimes observed that recognition performance for simple context-independent subword-unit recognizers can be degraded by RASTA. However, we have seen that RASTA has worked well when there is some explicit modeling of the effects of context from the past. This is so because RASTA processing increases the dependence on this context because of its filtering process.

Similarly, incorporating more advanced auditory models may not be fruitful without a similarly developed statistical pattern classification system that can take advantage of the auditory features. For this reason, we are now working on a statistical model of speech that is intrinsically perceptual. As with the human perceptual system, our new model (called the Stochastic Perceptual Auditory-event-based Model, or SPAM) focuses on novel events, such as major changes in the speech spectrum. As of this writing we have developed the basic mathematical theory [7], but we have not yet developed a practical system. The major idea of this development is that we model speech as a succession of auditory events, or detected novelties, which the perceptual system must disambiguate. These events are connected by periods in which the acoustics don't change much. The major difference from classical statistical speech models is that the system is not trained to discriminate between fine differences in steady-state regions, but rather to discriminate between regions of strong

transition that correspond to these auditory events. For instance, in the syllable "ma", the transition between the "m" and the "a" sound would receive more emphasis in the training of the statistical system than the more constant portions of the two sounds. This is very different from the modeling done in current recognizers, in which more emphasis is placed on the longer and more constant middle parts of the two sounds. We believe that this may be a bias of human speech perception as well, and that it should be a better match to auditory features that have the potential to improve acoustical robustness for the overall system.

Figure 1 shows the general structure of speech recognition in terms of some of the acoustical sources of error and the broad kinds of solutions that are being developed.

4 Conclusions

As human listeners, we tend to take for granted how well we can understand speech in the presence of strong acoustic interference. As speech system designers, however, we soon learn that "adverse conditions" for our recognizers correspond to "normal conditions" for people.

So far, our algorithms have taken advantage of an extremely simple property of human hearing to improve performance in the presence of noise and linear filtering. In the study reported here, we cut the error for a moderately noisy case by roughly a factor of three using these methods. However, the resulting error rate for the noisy case was still four times larger than it had been without added noise. We continue to derive inspiration from human perception in the development of speech recognition theory and systems. In particular, we are now working on the development of a statistical system that will focus modeling power on those acoustical segments that are critical to speech perception. If this research direction (or some other approach to acoustical robustness) proves successful, we can expect significant improvements in performance of commercial systems in future years, since practical systems operate in less controlled acoustic conditions than are found in the laboratory.

5 Acknowledgements

Much of the work described here has been done in conjunction with Hynek Hermansky of the Oregon Graduate Institute. H. Guenter Hirsch of the U. of Aachen also collaborated on some earlier work. Joachim Koehler and Grace Tong have worked on the experiments. The newer work is also benefiting from discussions with Steve Greenberg of UC Berkeley and Herveè Bourlard of ICSI. Over the last year these experiments have been partially sponsored (at UC Berkeley) under the Joint Services Electronics Program by contract number F49620-93-C-0014. We also acknowledge continuing support from the International Computer Science Institute.

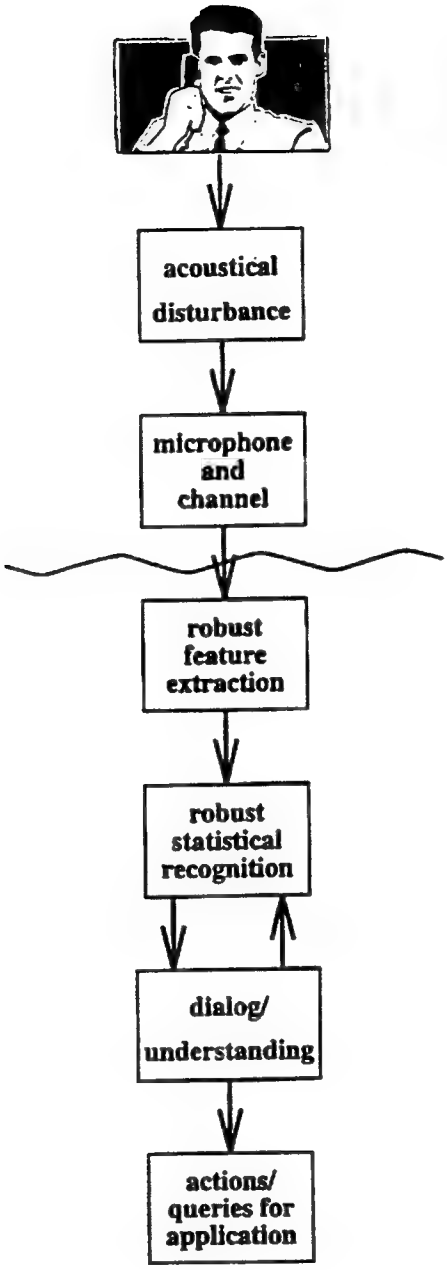


Figure 1: A simple block diagram of the speech recognition process. The blocks above the wavy line illustrate the steps in generating the signal that is received at the speech recognition system. This breakdown is chosen to show some of the sources of degradation in the received speech signal. The blocks below the wavy line are the major recognizer components, chosen to show the different pieces of the technology that can potentially be used to correct for the nonideal conditions that are created by realistic acoustical situations. This paper primarily refers to the feature extraction block, although we mention the statistical recognition block to some extent.

References

[1] S.F. Boll, Suppression of acoustic noise in speech using spectral subtraction, *IEEE ASSP-27*, pp. 113-120, Apr. 1979.

[2] H.G. Hirsch, Estimation of noise spectrum and its application to SNR-estimation and speech enhancement, *Technical Report TR-93-012*, ICSI, 1993

[3] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP) *Proc. EUROSPEECH '91*, pp. 1367-1370, Genova, 1991.

[4] G. Green, *Temporal Aspects of Audition* PhD Thesis, Oxford, 1976.

[5] H. Hermansky, N. Morgan, H.G. Hirsch, Recognition of speech in additive and convolutional noise based on RASTA spectral processing, *IEEE Proc. ICASSP'93*, pp. 83-86, 1993

[6] N. Morgan and H. Hermansky, RASTA extensions: Robustness to additive and convolutional noise , *Proc. Workshop on Speech Processing in Adverse Conditions*, Cannes, France, November 1992

[7] N. Morgan, H. Bourlard, S. Greenberg, and H. Hermansky, Stochastic Perceptual Auditory-event-based Models for Speech Recognition, *Intl Conf on Spoken Language Processing*, Yokohama, Japan, 1994, In Press

[8] T. Stockham, T. Cannon, and R. Ingebretsen: Blind Deconvolution Through Digital Signal Processing, *Proc. IEEE,* Vol 63, pp. 678-692, April 1975.

[9] Q. Summerfield, A. Sidwell and T. Nelson: Auditory enhancement of changes in spectral amplitude, *J.Acoust. Soc. Am. 81, (3)*, pp. 700-708, March 1987.

[10] P. Woodland, and S. Young, The HTK Tied-State Continuous Speech Recognizer, Eurospeech '93, pp. 2207-2210, 1993

Learning Complex Boolean Functions: Algorithms and Applications

Arlindo L. Oliveira and **Alberto Sangiovanni-Vincentelli**
Dept of EECS
UC Berkeley
Berkeley CA 94720

Abstract

The most commonly used neural network models are not well suited to direct digital implementations because each node needs to perform a large number of operations between floating point values. Fortunately, the ability to learn from examples and to generalize is not restricted to networks of this type. Indeed, networks where each node implements a simple Boolean function (Boolean networks) can be designed in such a way as to exhibit similar properties. Two algorithms that generate Boolean networks from examples are presented. The results show that these algorithms generalize very well in a class of problems that accept compact Boolean network descriptions. The techniques described are general and can be applied to tasks that are not known to have that characteristic. Two examples of applications are presented: image reconstruction and hand-written character recognition.

1 Introduction

The main objective of this research is the design of algorithms for empirical learning that generate networks suitable for digital implementations. Although threshold gate networks can be implemented using standard digital technologies, for many applications this approach is expensive and inefficient. Pulse stream modulation [Murray and Smith, 1988] is one possible approach, but is limited to a relatively small number of neurons and becomes slow if high precision is required. Dedicated

boards based on DSP processors can achieve very high performance and are very flexible but may be too expensive for some applications.

The algorithms described in this paper accept as input a training set and generate networks where each node implements a relatively simple Boolean function. Such networks will be called Boolean networks. Many applications can benefit from such an approach because the speed and compactness of digital implementations is still unmatched by its analog counterparts. Additionally, many alternatives are available to designers that want to implement Boolean networks, from full-custom design to field programmable gate arrays. This makes the digital alternative more cost effective than solutions based on analog designs.

Occam's razor [Blumer *et al.*, 1987; Rissanen, 1986] provides the theoretical foundation for the development of algorithms that can be used to obtain Boolean networks that generalize well. According to this paradigm, simpler explanations for the available data have higher predictive power. The induction problem can therefore be posed as an optimization problem: **given a labeled training set, derive the less complex Boolean network that is consistent[1] with the training set.**

Occam's razor, however, doesn't help in the choice of the particular way of measuring complexity that should be used. In general, different types of problems may require different complexity measures. The algorithms described in section 3.1 and 3.2 are greedy algorithms that aim at minimizing one specific complexity measure: the size of the overall network. Although this particular way of measuring complexity may prove inappropriate in some cases, we believe the approach proposed can be generalized and used with minor modifications in many other tasks. The problem of finding the smallest Boolean network consistent with the training set is NP-hard [Garey and Johnson, 1979] and cannot be solved exactly in most cases. Heuristic approaches like the ones described are therefore required.

2 Definitions

We consider the problem of supervised learning in an attribute based description language. The attributes (input variables) are assumed to be Boolean and every exemplar in the training set is labeled with a value that describes its class. Both algorithms try to maximize the mutual information between the network output and these labels.

Let variable X take the values $\{x_1, x_2, ...x_n\}$ with probabilities $p(x_1), p(x_2)...p(x_n)$. The entropy of X is given by $H(X) = -\sum_j p(x_j) \log p(x_j)$ and is a measure of the uncertainty about the value of X. The uncertainty about the value of X when the value of another variable Y is known is given by $H(X|Y) = -\sum_i p(y_i) \sum_j p(x_j|y_i) \log p(x_j|y_i)$.

The amount by which the uncertainty of X is reduced when the value of variable Y is known, $I(Y,X) = H(X) - H(X|Y)$ is called the mutual information between Y and X. In this context, Y will be a variable defined by the output of one or more nodes in the network and X will be the target value specified in the training set.

[1] Up to some specified level.

3 Algorithms

3.1 Muesli - An algorithm for the design of multi-level logic networks

This algorithm derives the Boolean network by performing gradient descent in the mutual information between a set of nodes and the target values specified by the labels in the training set.

In the pseudo code description of the algorithm given in figure 1, the function $\mathcal{I}(S)$ computes the mutual information between the nodes in S (viewed as a multi-valued variable) and the target output.

```
muesli(nlist) {
    nlist ← sort_nlist_by_I(nlist,1);
    sup ← 2;
    while (not_done(nlist) ∧ sup < max_sup) {
        act ← 0;
        do {
            act++;
            success ← improve_mi(act, nlist, sup);
        } while (success = FALSE ∧ act < max_act);
        if (success = TRUE) {
            sup ← 2;
            while (success = TRUE)
                success ← improve_mi(act, nlist, sup);
        }
        else sup++;
    }
}

improve_mi(act, nlist, sup) {
    nlist ← sort_nlist_by_I(nlist, act);
    f ← best_function(nlist, act, sup);
    if (I(nlist[1:act-1] ∪ f) > I(nlist[1:act])) {
        nlist ← nlist ∪ f;
        return(TRUE);
    }
    else return(FALSE);
}
```

Figure 1: Pseudo-code for the *Muesli* algorithm.

The algorithm works by keeping a list of candidate nodes, *nlist*, that initially contains only the primary inputs. The *act* variable selects which node in *nlist* is active. Initially, *act* is set to 1 and the node that provides more information about the output is selected as the active node. Function *improve_mi*() tries to combine the active node with other nodes as to increase the mutual information.

Except for very simple functions, a point will be reached where no further improve-

ments can be made for the single most *informative* node. The value of *act* is then increased (up to a pre-specified maximum) and *improve_mi* is again called to select auxiliary features using other nodes in *nlist* as the active node. If this fails, the value of *sup* (size of the support of each selected function) is increased until no further improvements are possible or the target is reached.

The function *sort_nlist_by_$\mathcal{I}$(nlist, act)* sorts the first *act* nodes in the list by decreasing value of the information they provide about the labels. More explicitly, the first node in the sorted list is the one that provides maximal information about the labels. The second node is the one that will provide more additional information after the first has been selected and so on.

Function *improve_mi*() calls *best_function(nlist, act, sup)* to select the Boolean function f that takes as inputs node *nlist[act]* plus *sup*−1 other nodes and maximizes $\mathcal{I}(nlist[1 : act-1] \cup f)$. When *sup* is larger than 2 it is unfeasible to search all $2^{2^{sup}}$ possible functions to select the desired one. However, given *sup* input variables, finding such a function is equivalent to selecting a partition[2] of the 2^{sup} points in the input space that maximizes a specific cost function. This partition is found using the Kernighan-Lin algorithm [Kernighan and Lin, 1970] for graph-partitioning.

Figure 2 exemplifies how the algorithm works when learning the simple Boolean function $f = ab + cde$ from a complete training set. In this example, the value of *sup* is always at 2. Therefore, only 2 input Boolean functions are generated.

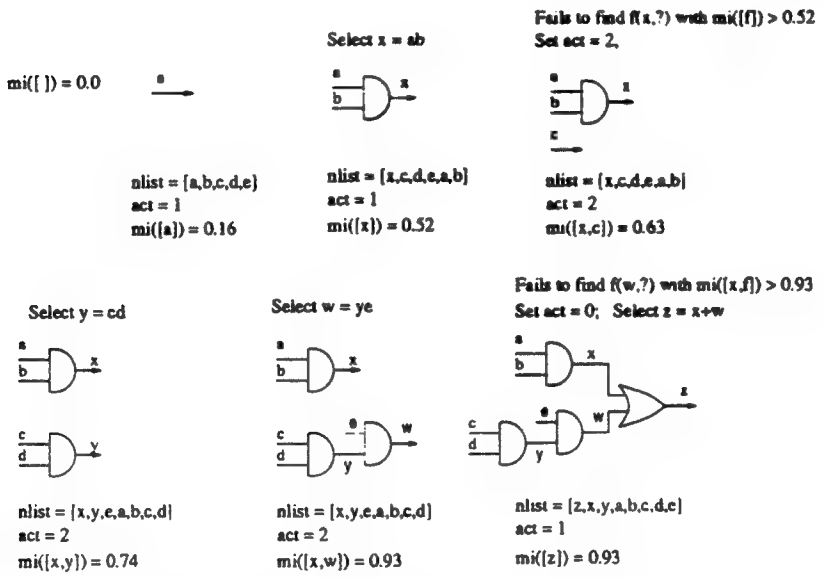


Figure 2: The *muesli* algorithm, illustrated

[2]A single output Boolean function is equivalent to a partition of the input space in two sets.

3.2 Fulfringe - a network generation algorithm based on decision trees

This algorithm uses binary decision trees [Quinlan, 1986] as the basic underlying representation. A binary decision tree is a rooted, directed, acyclic graph, where each terminal node (a node with no outgoing edges) is labeled with one of the possible output labels and each non-terminal node has exactly two outgoing edges labeled 0 and 1. Each non-terminal node is also labeled with the name of the attribute that is tested at that node. A decision tree can be used to classify a particular example by starting at the root node and taking, until a terminal is reached, the edge labeled with the value of the attribute tested at the current node.

Decision trees are usually built in a greedy way. At each step, the algorithm greedily selects the attribute to be tested as the one that provides maximal information about the label of the examples that reached that node in the decision tree. It then recurs after splitting these examples according to the value of the tested attribute.

Fulfringe works by identifying patterns near the fringes of the decision tree and using them to build new features. The idea was first proposed in [Pagallo and Haussler, 1990].

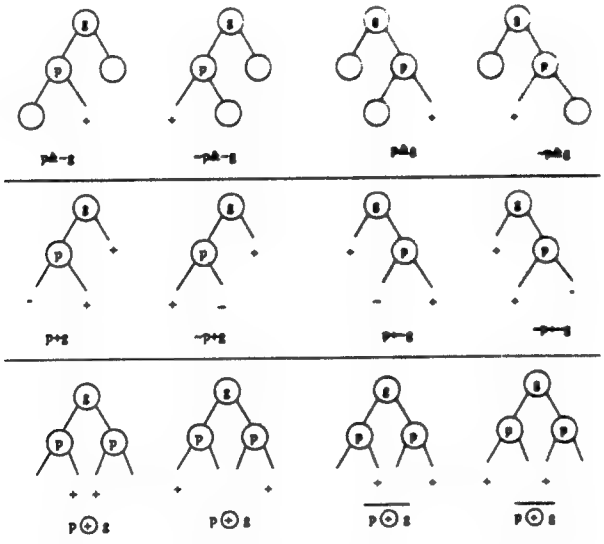


Figure 3: Fringe patterns identified by *fulfringe*

Figure 3 shows the patterns that *fulfringe* identifies. *Defringe*, proposed in [Yang *et al.*, 1991], identifies the patterns shown in the first two rows. These patterns correspond to 8 Boolean functions of 2 variables. Since there are only 10 distinct Boolean functions that depend on two variables[3], it is natural to add the patterns in the third row and identify all possible functions of 2 variables. As in *defringe* and *fringe*, these new composite features are added (if they have not yet been generated) to the list of available features and a new decision tree is built. The

[3]The remaining 6 functions of 2 variables depend on only one or none of the variables.

process is iterated until a decision tree with only one decision node is built. The attribute tested at this node is a complex feature and can be viewed as the output of a Boolean network that matches the training set data.

3.3 Encoding multivalued outputs

Both *muesli* and *fulfringe* generate Boolean networks with a single binary valued output. When the target label can have more than 2 values, some encoding must be used. The prefered solution is to encode the outputs using an error correcting code [Dietterich and Bakiri, 1991]. This approach preserves most of the compactness of a digital encoding while beeing much less sensitive to errors in one of the output variables. Additionally, the Hamming distance between an observed output and the closest valid codeword gives a measure of the certainty of the classification. This can be used to our advantage in problems where a failure to classify is less serious than the output of a wrong classification

4 Performance evaluation

To evaluate the algorithms, we selected a set of 11 functions of variable complexity. A complete description of these functions can be found in [Oliveira, 1994]. The first 6 functions were proposed as test cases in [Pagallo and Haussler, 1990] and accept compact *disjoint normal form* descriptions. The remaining ones accept compact multi-level representations but have large two level descriptions. The algorithms described in sections 3.1 and 3.2 were compared with the *cascade-correlation* algorithm [Fahlman and Lebiere, 1990] and a standard decision tree algorithm analog to ID3 [Quinlan, 1986]. As in [Pagallo and Haussler, 1990], the number of examples in the training set was selected to be equal to $\frac{1}{\epsilon}$ times the description length of the function under a fixed encoding scheme, where ϵ was set equal to 0.1. For each function, 5 training sets were randomly selected. The average accuracy for the 5 runs in an independent set of 4000 examples is listed in table 1.

Table 1: Accuracy of the four algorithms.

Function	# inputs	# examples	Accuracy			
			muesli	fulfringe	ID3	CasCor
dnf1	80	3292	99.91	99.98	82.09	75.38
dnf2	40	2185	99.28	98.89	88.84	73.11
dnf3	32	1650	99.94	100.00	89.98	79.19
dnf4	64	2640	100.00	100.00	72.61	58.41
xor4_16	16	1200	98.45	100.00	75.20	99.91
xor5_32	32	4000	60.16	100.00	51.41	99.97
sm12	12	1540	99.90	100.00	99.81	98.98
sm18	18	2720	100.00	99.92	91.48	91.30
str18	18	2720	100.00	100.00	94.55	92.57
str27	27	4160	98.64	99.35	94.24	93.90
carry8	16	2017	99.50	98.71	96.70	99.22
Average			95.97	99.71	85.35	87.45

The results show that the performance of *muesli* and *fulfringe* is consistently su-

perior to the other two algorithms. *Muesli* performs poorly in examples that have many *xor* functions, due the greedy nature of the algorithm. In particular, *muesli* failed to find a solution in the alloted time for 4 of the 5 runs of *xor5_32* and found the exact solution in only one of the runs.

ID3 was the fastest of the algorithms and Cascade-Correlation the slowest. *Fulfringe* and *muesli* exhibited similar running times for these tasks. We observed, however, that for larger problems the runtime for *fulfringe* becomes prohibitively high and *muesli* is comparatively much faster.

5 Applications

To evaluate the techniques described in real problems, experiments were performed in two domains: noisy image reconstruction and handwritten character recognition. The main objective was to investigate whether the approach is applicable to problems that are not known to accept a compact Boolean network representation. The outputs were encoded using a 15 bit Hadamard error correcting code.

5.1 Image reconstruction

The speed required by applications in image processing makes it a very interesting field for this type of approach. In this experiment, 16 level gray scale images were corrupted by random noise by switching each bit with 5% probability. Samples of this image were used to train a network in the reconstruction of the original image. The training set consisted of 5x5 pixel regions of corrupted images (100 binary variables per sample) labeled with the value of the center pixel. Figure 4 shows a detail of the reconstruction performed in an independent test image by the network obtained using *fulfringe*.

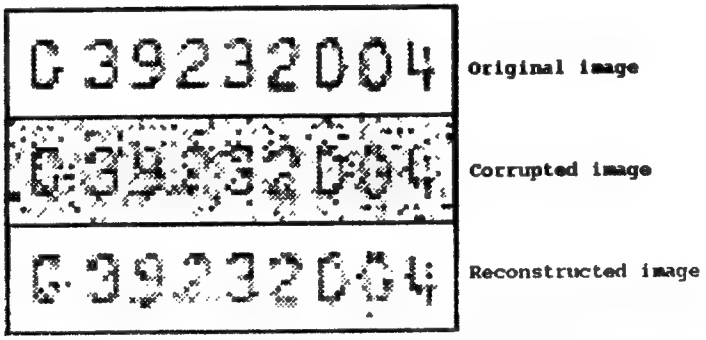


Figure 4: Image reconstruction experiment

5.2 Handwritten character recognition

The NIST database of handwritten characters was used for this task. Individually segmented digits were normalized to a 16 by 16 binary grid. A set of 53629 digits was used for training and the resulting network was tested in a different set of 52467

digits. Training was performed using *muesli*. The algorithm was stopped after a prespecified time (48 hours on a DECstation 5000/260) ellapsed The resulting network was placed and routed using the TimberWolf [Sechen and Sangiovanni-Vincentelli, 1986] package and occupies an area of 78.8 sq. mm. using 0.8μ technology.

The accuracy on the test set was 93.9%. This value compares well with the performance obtained by alternative approaches that use a similarly sized training set and little domain knowledge, but falls short of the best results published so far. Ongoing research on this problem is concentrated on the use of domain knowledge to restrict the search for compact networks and speed up the training.

Acknowledgements

This work was supported by Joint Services Electronics Program grant F49620-93-C-0014.

References

[Blumer *et al.*, 1987] A. Blumer, A. Ehrenfeucht, D. Haussler and M. Warmuth. Occam's razor. *Information Processing Letters*, 24:377–380, 1987.

[Dietterich and Bakiri, 1991] T. G. Dietterich and G. Bakiri. Error-correcting output codes: A general method for improving multiclass inductive learning programs. In *Proceedings of the Ninth National Conference on Artificial Intelligence (AAAI-91)*, pages 572–577. AAAI Press, 1991.

[Fahlman and Lebiere, 1990] S.E. Fahlman and C. Lebiere. The cascade-correlation learning architecture. In D.S. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2, pages 524–532, San Mateo, 1990. Morgan Kaufmann.

[Garey and Johnson, 1979] M.R. Garey and D.S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Freeman, New York, 1979.

[Kernighan and Lin, 1970] B. W. Kernighan and S. Lin. An efficient heuristic procedure for partitioning graphs. *The Bell System Technical Journal*, pages 291–307, February 1970.

[Murray and Smith, 1988] Alan F. Murray and Anthony V. W. Smith. Asynchronous vlsi neural networks using pulse-stream arithmetic. *IEEE Journal of Solid-State Circuits*, 23:3:688–697, 1988.

[Oliveira, 1994] Arlindo L. Oliveira. *Inductive Learning by Selection of Minimal Representations*. PhD thesis, UC Berkeley, 1994. In preparation.

[Pagallo and Haussler, 1990] G. Pagallo and D. Haussler. Boolean feature discovery in empirical learning. *Machine Learning*, 1, 1990.

[Quinlan, 1986] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

[Rissanen, 1986] J. Rissanen. Stochastic complexity and modeling. *Annals of Statistics*, 14:1080–1100, 1986.

[Sechen and Sangiovanni-Vincentelli, 1986]
Carl Sechen and Alberto Sangiovanni-Vincentelli. TimberWolf3.2: A new standard cell placement and global routing package. In *Proceedings of the 23rd Design Automation Conference*, pages 432–439, 1986.

[Yang *et al.*, 1991] D. S. Yang, L. Rendell, and G. Blix. Fringe-like feature construction: A comparative study and a unifying scheme. In *Proceedings of the Eight International Conference in Machine Learning*, pages 223–227, San Mateo, 1991. Morgan Kaufmann.

Combating Additive Noise and Spectral Distortion in Speech Recognition Systems with JAH-RASTA

Grace C.H. Tong

Submitted in partial fulfillment of the requirements of the degree
Master of Science in Electrical Engineering

Abstract

Speech recognizers often operate in an adverse environment: acoustic ambient noise, spectral distortions, reverberation and other environmental factors all cause severe degradation in the recognition performance. In this report, we discuss a front-end technique, called Jah-RASTA, to combat noise and improve the robustness of speech recognizers. Jah-RASTA processing uses bandpass filtering of temporal trajectories of non-linearly transformed critical band spectrum to simultaneously reduce additive noise and spectral distortion. However, the optimal form of the nonlinear transform used by Jah-RASTA is a function of the noise level - a time varying quantity. This introduces a new source of variability into the speech recognition system, and hurts recognition performance. To compensate for this new source of variability, a spectral mapping approach has been developed. The method shows improved robustness and is computationally efficient as well.

Abstract

The calibration of defocus distance and exposure time in lithographic equipment for integrated circuits fabrication is currently performed manually. An automated approach promises better consistency and reproducibility at a lower cost. The two critical parameters that determine the performance of a lithographic stepper are the defocus distance and the exposure time. Currently, the optimal settings are selected after examining a pattern that has been projected several times across one wafer. Each projection is done under a different combination of exposure time and defocus. The "best" pattern is chosen by an experienced operator, who looks for the image that appears to be the sharpest, having the most vertical sidewalls, and whose critical dimensions are the closest to those of the desired pattern. The focus and exposure settings corresponding to this image are then selected as the settings to use. This, for example, is done when choosing the best exposure and identifying current focus in using a SMILE or Bossung plot. This calibration procedure has to be repeated periodically since the stepper, the light source and the chemicals tend to age. Calibration is also necessary whenever maintenance is performed, or whenever the machine is configured for the patterning of a new layer.

In this project we applied a two dimensional pattern recognition network which was trained to choose the "best" developed image. We collected a database of digitized optical calibration images generated on our stepper and tagged with a qualification code supplied by a human expert. A feed forward network was trained using the backpropagation training algorithm to recognize key aspects of the patterns exposed under different stepper settings. We used image processing techniques (such as edge extraction and convolution) to pre-process the data before it

Vertical-cavity Laser Diodes Fabricated by Phase-locked Epitaxy

by

Jeffrey David Walker

B.S. (University of California at Berkeley) 1987
M.S. (University of California at Berkeley) 1990

A dissertion submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering — Electrical Engineering
and Computer Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY of CALIFORNIA at BERKELEY

Committee in charge:

Professor John Stephen Smith, Chair
Professor John R. Whinnery
Professor Herbert Steiner

1993

Abstract

Vertical-cavity Laser Diodes Fabricated by

Phase-locked Epitaxy

by

Jeffrey David Walker

Doctor of Philosophy in Electrical Engineering and Computer Sciences

University of California at Berkeley

Professor John Stephen Smith, Chair

The fabrication of vertical-cavity surface-emitting laser diodes (VCSELs) has challenged the capabilities of conventional thin film growth techniques such as MBE because of the stringent requirements on layer thickness and interface flatness required to produce high reflectivity AlGaAs Bragg reflectors. This dissertation presents a new growth technique for the fabrication of VCSELs that is based on phase-locked epitaxy (PLE) and that addresses the problems associated with the growth of these structures. The techniques presented here are the first to extend PLE toward the fabrication of precision macroscopic structures such as VCSELs.

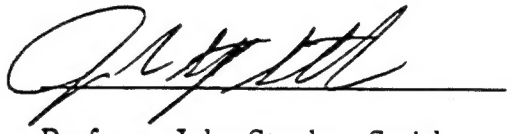
Exerimental and theoretical analysis show that PLE-grown Bragg reflectors have a maximum error in layer periodicity of 1%, and a maximum loss due to optical scattering of 0.01% per interface. In addition, 1.5% uniformity in layer thickness across a 50 mm wafer is achievable.

Because the PLE growth technique solves the thin-film growth problems that have hampered VCSEL development, it has been possible to fabricate extremely high quality lasers. The lasers presented here were the first to be fabricated using an in situ feedback technique to control layer thickness. They were also the first VCSELs to lase in the 10 mW CW power range, the first to utilize lower aluminum content mirrors to control series resistance, and the first with low threshold voltages

(1.6 V). Furthermore, this growth technique is capable of growing high quality VCSEL wafers within tight tolerances and with near 100% yields on both single-wafer and wafer-to-wafer scales.

In addition to results from VCSELs fabricated by PLE, a new type of integrable device called a deformable Fabry-Perot (DFP) cavity is proposed. The DFP has potential applications for broadly-tunable surface-normal detectors and lasers for wavelength division multiplexing. Theoretical analysis indicates 50 nm wavelength tunability with a 10 V applied bias.

Finally, a new type of AlGaAs multiple quantum well 5-20 μm mid-infrared detector is proposed. The device uses modification of the conduction band wave functions to allow for the absorption of normal incidence light. Theoretical analysis shows the ability to achieve 1000 cm^{-1} absorption.



Professor John Stephen Smith

Committee Chairman

Hot-Carrier Currents of SOI MOSFETs

Hsing-jen Wann, Joe King, Jian Chen, Ping K. Ko, and Chenming Hu
Department of Electrical Engineering and Computer Sciences
University of California at Berkeley, CA 94720

MOSFETs built on the SOI structure exhibit superior short channel behaviors over the bulk MOSFETs[1]. They also have other advantages such as reduction of the junction capacitance, radiation hardness and ease for device isolation. The SOI MOSFET is a promising candidate for future device scaling. The hot-carrier effect that increases with device miniaturization is another important device scaling constraint that has to be considered for the SOI MOSFET. The hot-carrier effect, which is usually monitored by the substrate current for the NMOSFET and the gate current for the PMOSFET for bulk devices, are closely related to the high channel electric field near the drain[2]. The quasi-two-dimensional (quasi-2D) model provides the link between the hot-carrier currents and the device design parameters for the bulk MOSFETs[3]. In this model the maximum channel field is $E_m = \frac{V_D - V_{DSAT}}{\ell}$ with the characteristic length $\ell = 0.22 t_{OX}^{1/3} X_j^{1/2}$.

The hot-carrier currents are given by:

$$I_{SUB} = 1.2(V_D - V_{DSAT}) I_D e^{-1.7\times10^6/E_m} \quad (1)$$ for the substrate current in NMOSFET[2], and

$$I_G = 0.5 \frac{I_{SUB} t_{OX}}{\lambda_r} \left(\frac{\lambda E_m}{\Phi_b}\right) P(E_{OX}) e^{-\Phi_b/E_m\lambda} \quad (2)$$ for the gate current in PMOSFET[4].

There were many attempts trying to apply the quasi-2D model to study the hot-carrier effect of SOI MOSFETs[5-7]. However the substrate current can not be measured directly, and the definition of the junction depth X_j is missing in thin-film SOI MOSFETs. In some works t_{Si}, the silicon film thickness has be used for X_j. The validity was not justified.

Fig.1 shows the simulated maximum channel electric field compared to what the bulk model predicts with the X_j replaced with t_{Si}. Such a substitution results in too strong the dependence on t_{Si} to the channel field. Fig.2 shows that the lateral doping gradient plays an important role in determining the high channel field. Note that in the bulk MOSFET, the dependence X_j of the channel field comes from two effects, the 2D effect and the lateral junction gradient effects. These two effects are strongly correlated because X_j and the lateral diffusion are formed during the same processing steps. However in SOI MOSFET such a correlation is missing. We need to refine the model by considering the 2D effect and the lateral doping gradient effect separately. This is done by solving the Poisson equation in the velocity saturation region with the lateral drain doping profile approximated by the exponential function around the concentration of 2×10^{18}[8]:

$$E_y = E_{SAT} e^{y/\ell} - \frac{qN_D\lambda}{\varepsilon_{Si}} e^{y-y_o/\lambda}, \qquad \int_{y_o}^{y_{s,d}} E_y dy = V_D - V_{DSAT} \quad (3)$$

The maximum channel field is found to be: $E_m = \frac{V_D - V_{DSAT}}{\ell} \cdot FRF$, with FRF shown in Fig.3, and ℓ is now the device characteristic length for the MOSFET with abrupt junctions. For the MOSFET with thinner t_{Si}, the channel field is larger due to stronger 2D effect. The field penetrates into the drain junction with higher concentration and is reduced more. Therefore the sensitivity on t_{Si} is weakened.

Fig.4 and 5 show the experimental data of hot-carrier body currents for SOI NMOSFETs. These devices have special layouts for the body contacts to collect the hot-carrier currents generated by impact ionization. Good agreements between the data and the model (2) are found using parameters in the captions. Fig.6 show the hot-electron current data for SOI PMOSFETs. The ratio I_{Body}/I_D does not increase much with thinner t_{OX} and t_{Si}. Since the ratio I_{Gate}/I_{Body} depends the E_{OX}, the device with thinner t_{OX} would have larger gate currents.

[1] K.Young, TED Feb 1989
[2] C.Hu, IEDM 1983
[3] P.Ko, IEDM 1980
[4] T.Ong et al., TED July 1990.
[5] J.Colinge, TED Oct. 1987
[6] J.Chen et al.,IEDM 1990
[7] L.Su et al., IEDM 1992
[8] H.Wann et al., VLSITSA 1993

ACKNOWLEDGEMENT This project is supported by SRC, ISTO/SDIO through ONR under contract number N00014-85-K-0603 and AFOSR/JSEP and SRC under contract number 93-DC-324.

0-7803-1346-1/93/$3.00 © 1993 IEEE

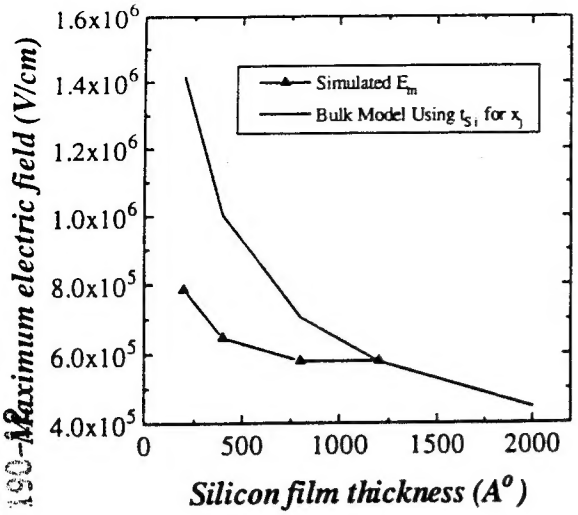


Fig.1 The maximum channel electric field, by using t_{Si} for X_j in the model for the bulk MOSFET, and by SPISCES simulation.

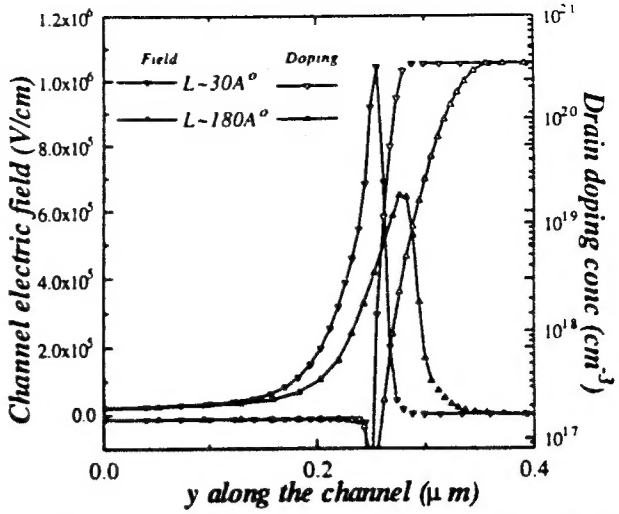


Fig.2 The channel field rises exponentially in the velocity saturation region as predicted by the quasi-2D model. The lateral doping gradient is important in determining E_m.

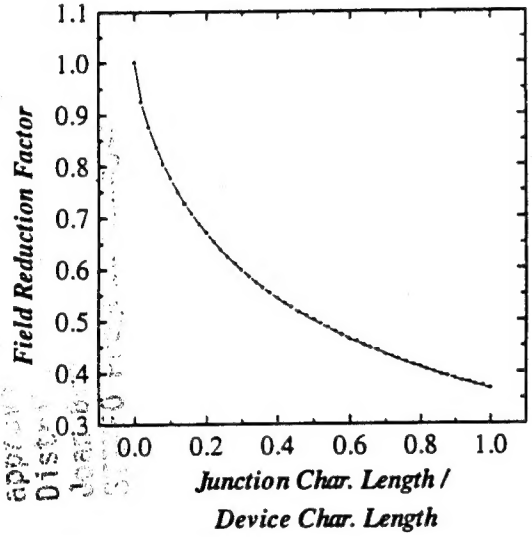


Fig.3 The "field reduction factor," defined as the reduction of E_m due to the finite drain junction gradient as opposed to an abrupt junctions, is a function of λ/ℓ .

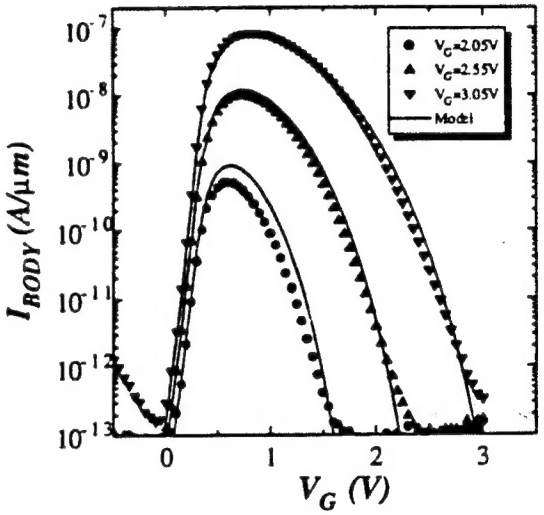


Fig.4 NMOSFET body currents for several V_G's compared with (2). The parameters used in the model are: t_{ox}=70A°, t_{Si}=560A°(both measured), and λ=130A°.

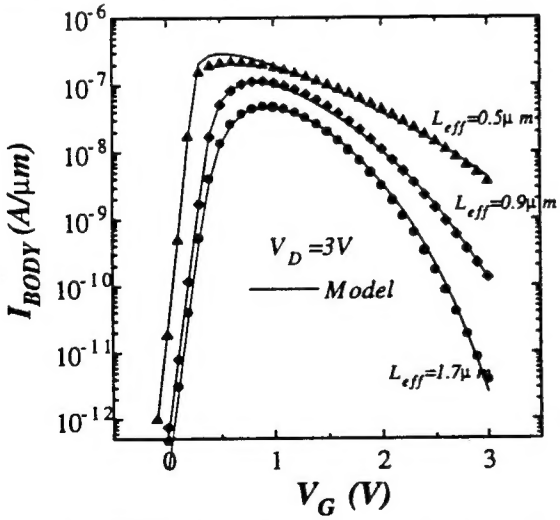


Fig.5 NMOSFET body currents for several channel lengths compared with (2). The parameters used in the model are: t_{ox}=90A°, t_{Si}=760A°(both measured), and λ=130A°.

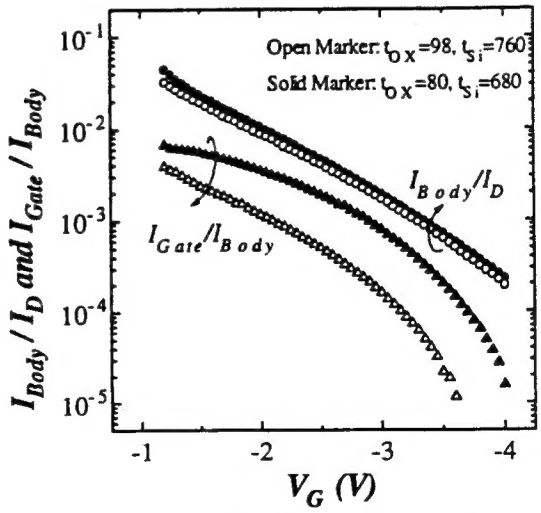


Fig.6 The measured PMOSFET body currents and gate currents.